



THE UNIVERSITY OF  
**SYDNEY**

PHD THESIS

---

# Vision Language Model for Medical Image Analysis

---

*Author:*

Yunyi LIU

*Supervisor:*

A/Prof. Luping ZHOU

*Co-Supervisor:*

A/Prof. Dong YUAN

*A thesis submitted in fulfilment of the requirements  
for the degree of Doctor of Philosophy*

School of Electrical and Computer Engineering

Faculty of Engineering

*The University of Sydney*

November 24, 2025



## *Authorship Attribution Statement*

Chapter 3 of this thesis is published as,

**Yunyi Liu\***, Zhanyu Wang, Dong Xu, Luping Zhou. “Q2ATransformer: Improving Medical VQA via an Answer Querying Decoder.” (Accepted by INFORMATION PROCESSING IN MEDICAL IMAGING, 2023. (Oral))

Chapter 4 of this thesis is under review as,

**Yunyi Liu\***, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. “SAT-RRG: Self-Adaptive Training for Radiology Report Generation Leveraging LLMs for Dynamic Token-Level Refinement.” (Submitted to CVPR 2026)

Chapter 5 of this thesis is published as,

**Yunyi Liu\***, Yingshu Li, Zhanyu Wang, Xinyu Liang, Lei Wang, Lingqiao Liu, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. “A Systematic Evaluation of GPT-4V’s Multimodal Capability for Medical Image Analysis.” (Meta-radiology 2024)

Chapter 6 of this thesis is published as,

**Yunyi Liu\***, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. “MRScore: Evaluating Radiology Report Generation with LLM-based Reward System.” arXiv preprint arXiv:2404.17778, 2024. (Early accepted by MICCAI 2024)

Chapter 7 of this thesis is published as,

**Yunyi Liu\***, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. “ReFINE: A Reward-Based Framework for Interpretable and Nuanced Evaluation of Radiology Report Generation.” (Accepted by AAI 2026)

In all these publications, I proposed the methods, conducted the experiments, and wrote the papers. My primary supervisor, Associate Professor Luping Zhou, is the corresponding author on all submissions.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Student Name: \_\_\_\_\_

Date: \_\_\_\_\_

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name:: \_\_\_\_\_

Date: \_\_\_\_\_

## *Generative AI Statement*

During the preparation of this thesis, the author used ChatGPT (developed by OpenAI) for the purposes of text enhancement. The use of this generative AI tool included paraphrasing, improving sentence structure, and correcting spelling and grammatical issues. The author confirms that where text was modified by generative AI, the content was carefully reviewed for possible errors, inaccuracies, and bias. The author takes full responsibility for the submitted thesis and affirms that the work is their own. The use of generative AI adhered to the parameters outlined in the University of Sydney's generative AI guide for researchers.



Abstract of thesis entitled

# **Vision Language Model for Medical Image Analysis**

Submitted by

**Yunyi LIU**

for the degree of Doctor of Philosophy

at The University of Sydney

in November, 2025

Vision-language models (VLMs) have emerged as powerful tools in medical image understanding by enabling joint reasoning over visual and textual information. Within this field, three core tasks have garnered significant attention: Visual Question Answering (VQA), Radiology Report Generation (RRG), and Visual Grounding (VG). VQA focuses on providing accurate answers to clinical questions grounded in medical images. RRG involves automatically generating coherent and clinically meaningful diagnostic reports. VG aims to align textual descriptions with relevant visual regions in the image, supporting explainability and localization.

## **Visual Question Answering in medical image**

Medical Visual Question Answering (VQA) involves interpreting clinically relevant image content via both close-ended and open-ended queries. While classification models excel at close-ended questions, they struggle with expressive answers required by open-ended ones. Generative models offer flexibility but often hallucinate, compromising clinical accuracy. To bridge this gap, we propose Q2ATransformer, a unified Transformer-based framework that treats both question types within a single architecture. It integrates a learnable set of candidate answer embeddings into a Transformer decoder, enabling dynamic, answer-aware attention over fused image-question features. This hybrid design maintains the efficiency and reliability of classification while capturing the semantic

richness of generative approaches. Experiments show Q2ATransformer achieves state-of-the-art results on VQA-RAD (79.19)

### **Radiology Report Generation in medical image**

Radiology Report Generation (RRG) aims to produce clinically accurate reports from medical images, bridging vision and language to support radiological workflows. Existing models rely heavily on cross-entropy (CE) loss, which enforces exact token-level matching but overlooks semantically valid alternatives and factual correctness. To overcome this, we propose SAT-RRG, a self-adaptive training framework that leverages large language models (LLMs) for token-level supervision without human annotations. SAT-RRG identifies semantically erroneous spans during training and applies two tailored losses: Confident Token Adaptive Loss (CTAL) to reinforce correct predictions, and Error Token Adaptive Penalty Loss (ETAPL) to penalize confident yet incorrect tokens. This dual strategy enables the model to focus on clinically meaningful corrections while maintaining fluent generation. Built upon a unified LLM backbone, SAT-RRG introduces no inference overhead. Experiments on MIMIC-CXR and IU-Xray show SAT-RRG outperforms prior methods in both factual accuracy and clinical quality.

### **Current large vision-language models(VLM) zero-shot and few-shot ability in medical image**

Recent advances in vision-language models (VLMs) like GPT-4V have shown strong performance on general multimodal tasks with minimal fine-tuning. However, their effectiveness in the medical domain, where detailed clinical reasoning and domain-specific knowledge are crucial, remains unclear. This work provides a systematic evaluation of GPT-4V's capabilities on three core medical image understanding tasks: radiology report generation, medical visual question answering (VQA), and visual grounding.

We design tailored prompts to elicit GPT-4V's responses in zero-shot and few-shot settings, and evaluate its performance through:

- **Quantitative metrics** for output quality,

- **Expert human evaluation** for clinical correctness,
- **Case studies** for qualitative insights.

Results show that GPT-4V performs well in report generation and VQA, producing clinically plausible outputs. However, it struggles with visual grounding, often failing to localize relevant regions. Additionally, we observe a gap between automatic metrics and human judgment, emphasizing the need for more clinically aligned evaluation methods in medical VLM research.

### **Radiology report evaluation metrics in medical image**

Evaluating the clinical quality of radiology reports remains a key challenge in automated report generation, as conventional metrics like BLEU, ROUGE, and CIDEr fail to capture semantic accuracy, factual consistency, and clinical relevance. We propose MRScore, a novel LLM-based evaluation metric tailored for radiology report generation. MRScore is trained on synthetic report pairs representing accepted and rejected outputs, generated via GPT-based models and filtered using radiologist-informed criteria. This enables the reward model to learn nuanced clinical distinctions and produce scores that better align with expert judgments. Extensive experiments show that MRScore achieves substantially higher correlation with human ratings than existing metrics, providing a scalable, clinically grounded solution for evaluating radiology report quality.

### **Radiology report evaluation metrics with explainable sub-score in medical image**

While automated radiology report generation (R2Gen) has seen substantial advancements in recent years, evaluating the quality of generated reports remains a complex and unresolved challenge. Existing metrics—such as BLEU, ROUGE, and CIDEr—often exhibit weak correlation with human judgment, particularly in identifying subtle but clinically critical deficiencies in content, structure, or factual accuracy. To bridge this gap, we propose **ReFINE**, a novel evaluation framework that trains a dedicated evaluation model using a **margin-based reward enforcement**

**loss.** Unlike scalar metrics, ReFINE decomposes report quality into multiple **fine-grained sub-scores**, each aligned with a user-defined evaluation criterion (e.g., factual correctness, clinical coverage, coherence), thus enhancing interpretability and diagnostic utility. Our framework leverages GPT-4 to generate diverse training samples by pairing high-quality (accepted) and low-quality (rejected) reports. These paired samples are then used to supervise a reward model, which learns to assign criterion-specific scores that reflect meaningful distinctions in report quality. The resulting **ReFINE Score** consists of both an overall quality assessment and interpretable sub-scores, enabling a detailed, criterion-aware evaluation of model outputs. We validate ReFINE across multiple experimental settings. Empirical results show that ReFINE aligns significantly better with human expert evaluations than conventional metrics and provides more reliable model selection signals. Its robustness is demonstrated across three expert-annotated datasets, encompassing chest X-rays and multimodal imaging reports spanning nine medical imaging modalities, as well as under two distinct scoring schemes. *ReFINE thus represents a step toward clinically meaningful, transparent, and fine-grained evaluation of radiology report generation systems.*

# **Vision Language Model for Medical Image Analysis**

by

**Yunyi LIU**

A Thesis Submitted in Fulfilment  
of the Requirements for the Degree of  
**Doctor of Philosophy**

at

University of Sydney  
November, 2025

COPYRIGHT ©2025, BY YUNYI LIU  
ALL RIGHTS RESERVED.

# Declaration

I, Yunyi LIU, declare that this thesis titled, "Vision Language Model for Medical Image Analysis", which is submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy, represents my own work except where due acknowledgement has been made. I further declared that it has not been previously included in a thesis, dissertation, or report submitted to this University or to any other institution for a degree, diploma, or other qualifications.

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

*For My Family*

## *Acknowledgements*

I extend my sincere gratitude to all individuals and organizations whose contributions were instrumental in the successful completion of my doctoral thesis.

First and foremost, I would like to express my heartfelt appreciation to my supervisor, **Prof. Luping Zhou**. Her selfless guidance and unwavering dedication throughout my PhD journey have been truly invaluable. I will always treasure the nights we stayed up late working on papers together — those moments are some of the most unforgettable of my life. Prof. Zhou's meticulous supervision of my academic development, along with her support of my progress, made a profound difference in my research journey. She is an exceptional mentor, and I feel incredibly fortunate to have had her guidance. My PhD experience has been deeply rewarding, and a significant part of that is thanks to having such an outstanding supervisor.

I want to thank my previous supervisor, Dong Xu, and also extend my gratitude to my auxiliary supervisors, **Prof. Lei Wang** from the University of Wollongong and **Prof. Lingqiao Liu** from the University of Adelaide, for their generous support and mentorship. Prof. Wang always provided critical insights at key moments, helping elevate my research to a higher level. Prof. Liu consistently offered valuable suggestions in the area of natural language processing, significantly enriching the depth and relevance of my work. Their contributions made my research both more impactful and more meaningful.

I would also like to thank the peers and friends who accompanied me on this research journey. In particular, I am grateful to my senior lab mate Dr. Zhanyu Wang, whose guidance was especially important when I first began. I also thank my fellow student, Mr. Yingshu Li, with whom I shared countless discussions on academic ideas. Along with many other colleagues and friends: Dr. Xinhui Liu, Dr. Xinyu Liang, Dr. Erjian Guo, Dr. Zhenghao Chen, Mr. Tong Chen, Mr. Ling Yang, Mr.

Lengkangping Ji, Ms. Yaqi Liu, Ms. Ying Zhou, and Ms. Yi Xia, they were by my side during some of the most formative years of my life, and I sincerely appreciate their support.

I am deeply thankful to my family for their unconditional love and support. My parents have always stood behind me, fully supporting my choices both emotionally and financially. As long as I was happy and committed, they would go out of their way to help me analyze how to pursue my goals more effectively. I feel incredibly lucky to have such devoted and encouraging parents. I also want to thank my dear younger sister, **Yulin Liu**, with whom I share a very close bond. She has always been a comforting presence during difficult times, and I am so grateful for the joy she brings me.

Finally, I must express my love and thanks to my adorable cat and dog, **A Fu** and **Yuan Bao**. During countless late-night experiments and moments of stress, it was their quiet companionship that brought me calm and happiness. I am truly thankful they chose to be part of my life.

In closing, I extend my heartfelt appreciation to everyone who contributed to my doctoral journey. Your support and encouragement have been vital in shaping both my academic path and personal growth. Thank you for being part of this important chapter in my life.

Yunyi LIU  
University of Sydney  
November 24, 2025



# List of Publications

## JOURNALS:

- [1] **Yunyi Liu\***, Yingshu Li, Zhanyu Wang, Xinyu Liang, Lei Wang, Lingqiao Liu, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. "A Systematic Evaluation of GPT-4V's Multimodal Capability for Medical Image Analysis." (*Meta-radiology 2024*)
- [2] Yingshu Li, **Yunyi Liu\***, Zhanyu Wang, Lingqiao Liu, Lei Wang and Luping Zhou. "S-RRG-Bench: Structured Radiology Report Generation with Fine-Grained Evaluation Framework" (*Meta-radiology 2025*)
- [3] Yingshu Li, **Yunyi Liu\***, Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. "Fine-Grained Disease-Aware Feature Alignment for Radiology Report Generation with Large Language Models." (Submitted to *IEEE Transactions on Medical Imaging*)
- [4] Yingshu Li, **Yunyi Liu\***, Lei Wang, Lingqiao Liu and Luping Zhou. "RadReason: Radiology Report Evaluation Metric with Reasons and Sub-Scores" (Submitted to *Artificial Intelligence in Medicine*)
- [5] Shaoyang Zhou, Yingshu Li, **Yunyi Liu\***, Lingqiao Liu, Lei Wang, and Luping Zhou. "A Review of Longitudinal Radiology Report Generation: Dataset Composition, Methods, and Performance Evaluation" (Submitted to *Medical Image Analysis*)

## CONFERENCES:

- [1] **Yunyi Liu\***, Zhanyu Wang, Dong Xu, Luping Zhou. "Q2ATransformer: Improving Medical VQA via an Answer Querying Decoder." (*INFORMATION PROCESSING IN MEDICAL IMAGING 2023*(Oral))
- [2] **Yunyi Liu\***, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. "MRScore: Evaluating Radiology Report Generation with LLM-based Reward System." (Early accepted by *MICCAI 2024*)
- [3] **Yunyi Liu\***, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. "ReFINE: A Reward-Based Framework for Interpretable and Nuanced Evaluation of Radiology Report Generation." (Accepted by *AAAI 2026*)
- [4] **Yunyi Liu\***, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. "SAT-RRG: Self-Adaptive Training for Radiology Report Generation Leveraging LLMs for Dynamic Token-Level Refinement." (Submitted to *CVPR 2026*)
- [5] Yingshu Li, Zhanyu Wang, **Yunyi Liu\***, Lei Wang, Lingqiao Liu, and Luping Zhou. "KARGEN: Knowledge-enhanced Automated Radiology Report Generation Using Large Language Models." (Early accepted by *MICCAI 2024*)
- [6] Erjian Guo, Zhen Zhao, Zicheng Wang, Tong Chen, **Yunyi Liu\***, and Luping Zhou. "DiN: Diffusion Model for Robust Medical VQA with Semantic Noisy Labels." (*CVPR 2025 (Accepted)*)
- [7] Yingshu Li, **Yunyi Liu\***, Zhenghao Chen, Tong Chen, Lingqiao Liu, Lei Wang, and Luping Zhou. "Seeing What Matters: Lesion-Aware High-Resolution Patch Discovery and Fusion for Chest X-ray Report Generation" (Submitted to *CVPR 2026*)

# Contents

|   |             |
|---|-------------|
| <b>Authorship Attribution Statement</b>   | <b>i</b>    |
| <b>Generative AI Statement</b>  | <b>i</b>    |
| <b>Abstract</b>   | <b>iii</b>  |
| <b>Declaration</b>  | <b>i</b>    |
| <b>Acknowledgements</b>   | <b>ii</b>   |
| <b>List of Publications</b>   | <b>v</b>    |
| <b>List of Figures</b>  | <b>xiii</b> |
| <b>List of Tables</b>   | <b>xix</b>  |
| <b>1 Introduction</b>   | <b>1</b>    |
| 1.1 Background and Problem Statement . . . . .  | 1           |
| 1.2 Challenges and Motivations . . . . .  | 4           |
| 1.2.1 Challenges . . . . .  | 4           |
| Challenge 1: Medical Image Visual Question An-<br>swering (VQA) . . . . .                                   | 4           |
| Challenge 2: Radiology Report Generation (RRG) .  | 5           |
| Challenge 3: Evaluating Large VLMs for Zero-shot<br>and Few-shot Medical Image Understand-<br>ing . . . . . | 6           |
| Challenge 4: Limitations of Traditional Radiology<br>Report Evaluation Metrics . . . . .                    | 7           |

|   |           |
|---|-----------|
| Challenge 5: Lack of Fine-Grained and Practical Evaluation in Existing Radiology Report Metrics . . . . . | 7         |
| 1.2.2 Motivations . . . . .   | 8         |
| Motivation 1: Towards Unified and Accurate Medical VQA . . . . .  | 8         |
| Motivation 2: Improving Radiology Report Generation with Self-Adaptive Feedback . . . . .                 | 9         |
| Motivation 3: Benchmarking GPT-4V in Medical Vision-Language Tasks . . . . .                              | 10        |
| Motivation 4: Developing Human-Aligned and Scalable Evaluation Metrics . . . . .                          | 11        |
| Motivation 5: Toward Fine-Grained, Interpretable, and Deployable Evaluation with ReFINE . . . . .         | 12        |
| 1.3 Thesis Contribution and Outline . . . . .   | 14        |
| <b>2 Literature Review</b>  | <b>17</b> |
| 2.1 Visual Question Answering . . . . .   | 18        |
| 2.2 Radiology Report Generation . . . . .   | 20        |
| 2.3 Visual Grounding . . . . .  | 21        |
| 2.4 Radiology Report Evaluation Metrics . . . . .   | 23        |
| <b>3 Q2ATransformer: Improving Medical VQA via an Answer Querying Decoder</b>                             | <b>27</b> |
| 3.1 Introduction . . . . .  | 28        |
| 3.2 Method . . . . .  | 30        |
| 3.2.1 Visual-Question Encoder . . . . .   | 32        |
| 3.2.2 Answer Querying Decoder . . . . .   | 35        |
| 3.2.3 Loss Function . . . . .   | 36        |
| 3.3 Experiments And Results . . . . .   | 37        |
| 3.3.1 Datasets . . . . .  | 37        |
| 3.3.2 Comparison with the State-of-the-Art Methods . . . . .  | 37        |
| 3.3.3 Ablation Study . . . . .  | 38        |
| 3.3.4 Qualitative Results . . . . .   | 41        |
| 3.3.5 Limitation and Discussion . . . . .   | 41        |

|          |   |           |
|----------|---|-----------|
| <b>4</b> | <b>SAT-RRG: Self-Adaptive Training for Radiology Report Generation Leveraging LLMs for Dynamic Token-Level Refinement</b> | <b>45</b> |
| 4.1      | Introduction . . . . .  | 46        |
| 4.2      | Related Work . . . . .  | 48        |
| 4.3      | Methodology . . . . .   | 49        |
| 4.3.1    | Report Generation . . . . .   | 50        |
| 4.3.2    | Dynamic Error Token Identification and Correction   | 52        |
| 4.3.3    | Error-Aware Training Objectives . . . . .   | 53        |
| 4.3.4    | Example Workflow Illustration . . . . .   | 56        |
| 4.4      | Experiments . . . . .   | 59        |
| 4.4.1    | Datasets. . . . .   | 59        |
|          | MIMIC-CXR . . . . .   | 59        |
|          | IU-Xray . . . . .   | 59        |
| 4.4.2    | Implementation Details. . . . .   | 60        |
| 4.4.3    | Evaluation Metrics. . . . .   | 60        |
| 4.4.4    | Main Results. . . . .   | 61        |
| 4.4.5    | Ablation Study. . . . .   | 62        |
|          | loss components . . . . .   | 62        |
|          | Clinical analysis results across all 14 defined categories . . . . .  | 63        |
|          | Qualitative Analysis . . . . .  | 64        |
| <b>5</b> | <b>A Systematic Evaluation of GPT-4V’s Multimodal Capability for Chest X-ray Image Analysis</b>                           | <b>65</b> |
| 5.1      | Introduction . . . . .  | 66        |
| 5.2      | Related Works . . . . .   | 68        |
| 5.3      | Methodology and Experiments . . . . .   | 69        |
| 5.3.1    | Evaluation Tasks . . . . .  | 70        |
|          | Radiology Report Generation . . . . .   | 70        |
|          | Visual Question Answering . . . . .   | 72        |
|          | Visual Grounding . . . . .  | 72        |
| 5.3.2    | Evaluation Method/Process . . . . .   | 73        |
|          | Radiology Report Generation . . . . .   | 73        |
|          | Medical Visual Question Answering . . . . .   | 75        |
|          | Medical Visual Grounding: . . . . .   | 77        |

|          |   |            |
|----------|---|------------|
| 5.3.3    | Evaluation Datasets . . . . .   | 78         |
| 5.3.4    | Evaluation Result . . . . .   | 79         |
|          | Quantitative Results . . . . .  | 79         |
|          | Human Evaluation Results . . . . .  | 83         |
|          | Case Study . . . . .  | 88         |
| <b>6</b> | <b>MRScore: Evaluating Radiology Report Generation with LLM-based Reward System</b>                             | <b>103</b> |
| 6.1      | Introduction . . . . .  | 104        |
| 6.2      | Related Work . . . . .  | 106        |
| 6.3      | Method . . . . .  | 108        |
| 6.3.1    | Scoring Criteria Design and Scoring Dataset Generation . . . . .  | 108        |
| 6.3.2    | LLM-based Reward Model . . . . .  | 111        |
| 6.3.3    | Loss Function . . . . .   | 113        |
| 6.4      | Experiments and Result . . . . .  | 113        |
| 6.4.1    | Evaluation Dataset . . . . .  | 113        |
| 6.4.2    | Experiment Result . . . . .   | 114        |
| <b>7</b> | <b>ReFINE: A Reward-Based Framework for Interpretable and Nuanced Evaluation of Radiology Report Generation</b> | <b>117</b> |
| 7.1      | Introduction . . . . .  | 118        |
| 7.2      | Related work for LLM-based Evaluation Metrics . . . . .   | 120        |
| 7.3      | Method . . . . .  | 121        |
| 7.3.1    | Scoring Data Generation Pipeline . . . . .  | 122        |
| 7.3.2    | Reward Model . . . . .  | 125        |
|          | Model Input . . . . .   | 125        |
|          | Multi-Reward Generator . . . . .  | 125        |
|          | Multi-Reward Learning. . . . .  | 127        |
| 7.4      | Margin Reward Enforcement(MRE) Loss Function . . . . .  | 127        |
| 7.4.1    | Individual Reward Loss $\mathcal{L}_{\text{ind}}$ . . . . .   | 128        |
| 7.4.2    | Total Reward Loss $\mathcal{L}_{\text{tot}}$ . . . . .  | 128        |
| 7.4.3    | Margin Reward Enforcement (MRE) Loss . . . . .  | 129        |
| 7.5      | Experiments and Result . . . . .  | 130        |
| 7.5.1    | Datasets . . . . .  | 130        |
| 7.5.2    | Performance on ReXVal Dataset . . . . .   | 131        |

|          |  |            |
|----------|--|------------|
| 7.5.3    | Performance on RaTE-Eval Dataset . . . . .       | 132        |
| 7.5.4    | Performance on Rad-100 Dataset . . . . .         | 132        |
| 7.5.5    | Ablation study . . . . .                         | 134        |
| 7.5.6    | Qualitative Analysis . . . . .                   | 135        |
| <b>8</b> | <b>Conclusion and Future Work</b>                | <b>137</b> |
| 8.1      | Conclusion . . . . .                             | 137        |
| 8.2      | Future Work . . . . .                            | 138        |
| <b>A</b> | <b>Chapter 4 Appendix</b>                        | <b>143</b> |
| A.1      | Prompt for Error Token Annotation . . . . .      | 143        |
| <b>B</b> | <b>Chaper 5 Appendix</b>                         | <b>147</b> |
| B.1      | Details of Prompt Settings . . . . .             | 147        |
| B.1.1    | Zero-shot prompt . . . . .                       | 147        |
| B.1.2    | Few-shot prompt . . . . .                        | 147        |
| <b>C</b> | <b>Chapter 7 Appendix</b>                        | <b>151</b> |
| C.1      | Labeled data prompt design and results . . . . . | 151        |
| C.2      | Labeled data . . . . .                           | 155        |
| C.3      | Scoring Dataset Prompt Samples . . . . .         | 160        |
|          | <b>Bibliography</b>                              | <b>165</b> |



# List of Figures

- 2.1 The general pipeline of a medical Visual Question Answering (VQA) model. It includes four key components: image feature extraction, question feature extraction, feature fusion, and the final answer generation. . . . . 17
- 2.2 Two common types of questions in medical Visual Question Answering (VQA). Left: Yes/No (Close-ended) question, where the model answers binary clinical queries. Right: What/How (Open-ended) question, where the model generates descriptive clinical answers. . . . . 18
- 3.1 Different paradigm designs for medical VQA models. (a) The closed-type paradigm formulates VQA as a classification problem, where a prediction head directly operates on the joint representation of the image and question to identify the correct answer class. (b) The open-type paradigm follows a generative modeling approach, where the fused image-question features interact sequentially with previously generated answer tokens to produce the next token via a text decoder. (c) Our semi-open framework introduces a decoder to learn candidate answer embeddings, which then engage with the fused image-question features to enhance the accuracy of answer class prediction. . . . . 31

|     |  |    |
|-----|--|----|
| 3.2 | Overview of the Q2ATransformer architecture. The input pair, consisting of a medical image and a corresponding question, is processed by the Visual-Question Encoder, which extracts and integrates features from both modalities. This encoder includes a Swin Transformer for visual feature extraction, a BERT model for question encoding, and a Cross-Modality Attention Network designed to effectively fuse the two feature types. The resulting fused representation is then passed into the Answer-Querying Decoder. In this stage, learnable candidate answer embeddings serve as queries in an attention mechanism, interacting with the fused features to refine their representations and assess the likelihood of each potential answer being correct. . . . . | 32 |
| 3.3 | Ablation study about different dimensions of answer embeddings. . . . .  | 41 |
| 3.4 | Example results from PathVQA dataset. . . . .  | 42 |
| 3.5 | Example results from VQA-RAD dataset. . . . .  | 43 |
| 3.6 | Examples of model attention focusing on informative regions in the image and question. The visualizations highlight how the model aligns visual and textual features to generate the corresponding answer. . . . .   | 43 |

|     |   |    |
|-----|---|----|
| 4.1 | <p>Overview of the SAT-RRG framework. It includes three components: (a) Report generation process, (b) Error token identification via LLM-based semantic comparison, and (c) Inference process. Tokens or modules with the same color represent the same entity across different stages. After the generator produces the initial predicted report, it is passed through the LLM again for feedback. At this stage, each token in the predicted report retains its original generation probability, since the model has not yet been updated by any loss. The LLM-based comparison only annotates which tokens are semantically incorrect. These annotations serve as token-level supervision targets: incorrect tokens are penalized and correct tokens reinforced via two complementary losses (ETAPL and CTAL). The original probabilities are used by the losses to propagate training signals back into the model. . . . .</p> | 51 |
| 4.2 | <p>Overview of our proposed training framework with token-level supervision. During training, the predicted report is dynamically compared to the reference report using the LLM’s semantic understanding to identify correct and erroneous tokens. Tokens are annotated with <b>green</b> for correct and <b>red</b> for incorrect cases via inline tags. The <b>Correct Token Augmentation Loss (CTAL)</b> encourages higher confidence in correct tokens by promoting gradient flattening, while the <b>Error Token Adaptive Penalty Loss (ETAPL)</b> penalizes overconfident incorrect predictions. This token-level feedback facilitates interpretable and adaptive optimization for higher-quality generation. . . . .</p>  | 53 |
| 4.3 | <p><b>Gradient dynamics of CTAL and ETAPL.</b> CTAL applies stronger gradients to low-confidence correct tokens, reinforcing their confidence during training. ETAPL imposes larger penalties on high-confidence incorrect tokens, thereby suppressing overconfident semantic errors. This complementary gradient behavior enables stable and semantically guided optimization. . . . .</p>   | 56 |

|     |  |    |
|-----|--|----|
| 4.4 | Comparison of error tokens in the generated report before and after TLS. The highlighted errors are marked in corresponding colors to show the corrections made. . . . .   | 60 |
| 5.1 | The main evaluation process, depicted in the figure, encompasses the data generation process in the upper left, the evaluation process in the upper right, and, in the lower sections, the definition of criteria, doctor evaluation, analysis of NLP metrics, and case analysis. . . . .  | 69 |
| 5.2 | Three multimodal medical imaging tasks we employ to evaluate GPT-4V's performance. . . . .   | 71 |
| 5.3 | R2Gen Prompt Examples. Three types of prompt settings are tested, including one zero-shot prompt and three few-shot prompts. . . . .   | 74 |
| 5.4 | VQA Prompt Example. The content between the double braces is replaced with specific questions. . . . .   | 92 |
| 5.5 | VG Prompt Example. The content between double braces is replaced with specific image width, height, and description text related to the image. . . . .   | 93 |
| 5.6 | An illustration of GPT-4V's capability in generating medical reports using our designed few-shot mixed-example prompt. The ground truth does not mention a medical device but one is visibly present in the image, marked by red arrows. GPT-4V demonstrates the ability to recognize and describe the medical device in the generated report. . . . . | 93 |
| 5.7 | Distributions of scores based on human evaluation, BLEU, CIDEr, and F1, respectively. . . . .  | 94 |
| 5.8 | Scatter graphs show the relationship between BLEU/CIDEr/F1 score and the human evaluation, respectively. . . . .   | 94 |
| 5.9 | An example of a ground truth report and a GPT-4V generated report. Key medical information in the reports is highlighted using different colors. . . . .   | 95 |

|      |  |     |
|------|--|-----|
| 5.10 | Human evaluation of the correctness of the answers generated by GPT-4V on a subset of VQA-RAD, in contrast to the evaluation based on the classification accuracy of the same dataset. . . . .   | 95  |
| 5.11 | Two examples showcase GPT-4V’s performance on open-end questions. . . . .  | 96  |
| 5.12 | R2Gen case using zero-shot prompt. GPT-4V can generate radiology reports without example reports and convey both normal and abnormal aspects. For better illustration, the key medical information in the reports is highlighted using different colors. . . . .   | 97  |
| 5.13 | R2Gen normal case. Key medical information in the reports is highlighted using different colors. GPT-4V is more likely to generate reports containing abnormality descriptions when the prompt consists of only abnormal examples. The text in red corresponds to descriptions of abnormal conditions. . . . . | 98  |
| 5.14 | R2Gen abnormal case. Key medical information in the reports is highlighted using different colors. GPT-4V is more likely to generate normal reports when the prompt consists of only normal examples. The text in red corresponds to descriptions of normal conditions. . . . .                                | 99  |
| 5.15 | VQA case examples for close-end questions. By few-shot prompts, GPT-4V could discern the question type and generate correct "yes" or "no" answers. . . . .   | 99  |
| 5.16 | VQA case examples for open-end questions. By using few-shot prompts, GPT-4V has the capacity to generate correct answers (left). However, it may sometimes overlook visual information and generate answers solely based on general medical knowledge (right). . . . .   | 100 |
| 5.17 | Visual Grounding Prompt case. The bounding boxes in red are predicted by GPT-4V, while the ground truth bounding boxes are in green. . . . .   | 101 |

|     |   |     |
|-----|---|-----|
| 6.1 | The upper section depicts the process of generating synthetic training data using GPT-based report sampling across different quality levels. The lower section illustrates the reward model training procedure via LoRA adaptation. Solid arrows denote steps executed during training, while dashed arrows indicate the inference phase. . . . .   | 109 |
| 6.2 | Correlations between varied evaluation metrics and radiologist scores . . . . .   | 116 |
| 7.1 | Training procedure of the ReFINE model. . . . .   | 120 |
| 7.2 | Overview of the framework for our model. The diagram is divided into four main parts: 1) <b>Data Generator</b> : Generates reports and corresponding scores based on specified criteria. 2) <b>Pairing Rule</b> : Demonstrates the scoring and pairing process using four criteria as an example. Reports are paired into "accepted" and "rejected" categories based on their total scores and margins. 3) <b>Training Process</b> : Utilizes paired data to train a reward model through a LoRA-based large language model (LLM) to optimize the MRE loss. 4) <b>Inference Process</b> : rate the predicted report by comparing it to the ground truth report, generating both sub-scores and total scores for evaluation. . . . . | 122 |
| 7.3 | An visual example of ReFINE from ReXVal Dataset. The highlighted sentences in reports and their corresponding scores share the same colors. . . . .   | 136 |
| B.1 | Zero-shot prompt. No additional information was provided to GPT-4V. . . . .   | 148 |
| B.2 | Example reports in prompts: Three pairs of different example reports in few-shot prompt settings. We added these example reports to few-shot prompts to help GPT-4V generate radiology reports. . . . .   | 149 |

# List of Tables

|     |  |    |
|-----|--|----|
| 2.1 | Traditional evaluation metrics for radiology report generation. . . . .  | 24 |
| 2.2 | LLM-based evaluation metrics for radiology report generation. . . . .  | 25 |
| 3.1 | Performance comparison of different methods. † and ‡ indicate the methods are classification-based(closed-type) or generation-based(open-type), respectively. . . . .  | 39 |
| 3.2 | Ablation Studies. BAN, SAN, and CMAN stand for Bilinear Attention Network [50], Stacked Attention Network [118] and ours Cross-modality Attention Network, respectively; Decoder refers to our Answer-Querying Decoder. . . . .  | 39 |
| 4.1 | Comparison on MIMIC-CXR and IU-Xray datasets(@B stands for BLUE). . . . .  | 57 |
| 4.2 | Evaluation of Clinic-related Metrics on MIMIC-CXR. . . . .   | 61 |
| 4.3 | Ablation study for loss components. . . . .  | 61 |
| 4.4 | Disease-wise comparison between our model and the baseline across multiple metrics. . . . .  | 63 |
| 4.5 | Clinical metric comparison on the MIMIC-CXR dataset. We evaluate a range of recent radiology report generation models using RadGraph F1 ( $RG_{F1}$ ), BERTScore (BERT), RadCliQ (CliQ), GREEN, and the composite RaTEScore(RaTE). Higher values indicate better performance except for CliQ. Best and second-best results are highlighted in <b>bold</b> and <u>underline</u> . EKAGen is shaded to indicate that it uses higher-resolution images ( $300 \times 300$ ) than the other methods. . . . . | 64 |
| 5.1 | Comparison on the MIMIC-CXR dataset. . . . .   | 80 |

|     |   |     |
|-----|---|-----|
| 5.2 | Evaluation of Clinical Efficacy on MIMIC-CXR dataset. . .   | 81  |
| 5.3 | Results on VQA-RAD benchmark . . . . .  | 82  |
| 5.4 | mIoU(%) results on MS-CXR benchmark. . . . .  | 83  |
| 5.5 | Evaluation of P-Value and Kendall’s Tau . . . . .   | 87  |
| 6.1 | Table for error types and design detail. . . . .  | 110 |
| 6.2 | Verification of GPT-4’s Capacity in Generating Human-like Ratings . . . . .   | 110 |
| 6.3 | Assessment of Human Correlations on Evaluation Dataset  | 115 |
| 6.4 | Human Correlations of MRScore Using Different LLM Backbones . . . . .   | 116 |
| 7.1 | Human Correlations of ReFINE Subscores on ReXVal Dataset  | 131 |
| 7.2 | Human Correlation Comparison of Evaluation Metrics on ReXVal Dataset. <i>Gray part indicates online methods which directly use online model API.</i> . . . . .  | 133 |
| 7.3 | Human Correlation on the Multimodal RaTE-Eval Dataset (Sentence-level correlation). † indicates values directly cited from the RaTEScore paper; others are reproduced. . .  | 133 |
| 7.4 | Human Correlation (with p-values) on RaTE-Eval (Sentence-level, 9 modalities) . . . . .   | 133 |
| 7.5 | Subscore Accuracies on Rad-100 dataset. ‘Imp. Cons.’ stands for Impression Consistency, Imp. Org.’ for Impression Organ, ‘Desc. Les.’ for Description of Lesion, ‘Clin. Hist.’ for Clinical History, ‘Comp.’ for Completeness, ‘Gram.’ for Grammar, and ‘Med. Term.’ for Medical Terminology. . . . . | 134 |
| 7.6 | Human Correlation on Rad-100 Dataset. @B = BLEU-4, @R = ROUGE-L, @M = METEOR, @C = CIDEr, @BS = BERTScore, @RG = RadGraphF1, @Semb = semb_score, @RC = RadCliQ-v1 . . . . .   | 134 |
| 7.7 | Ablation Study of LLM Backbones on ReXVal Dataset . . .   | 135 |
| 7.8 | Ablation Study on Different Loss Terms . . . . .  | 135 |
| 7.9 | Varying $\lambda$ Values Ablation . . . . .   | 135 |

# Chapter 1

## Introduction

### 1.1 Background and Problem Statement

Vision-language models (VLMs) are a class of multimodal systems designed to jointly process and reason over visual and textual inputs. By integrating image and language understanding into a unified framework, VLMs enable a range of capabilities such as image captioning, visual question answering, and grounded reasoning. Recent large-scale VLMs, such as CLIP, Flamingo, and GPT-4V, have demonstrated strong zero-shot and few-shot performance across diverse general-domain tasks without requiring extensive task-specific fine-tuning.

In the medical domain, VLMs hold great promise for automating complex image interpretation tasks that traditionally require clinical expertise. Among these, three core applications have emerged as benchmarks for evaluating VLM capability: medical visual question answering (VQA), radiology report generation (RRG), and visual grounding (VG). Medical VQA involves answering clinically relevant questions based on medical images; RRG focuses on generating accurate, coherent diagnostic narratives; and VG targets aligning text descriptions with specific anatomical regions in medical scans. With the emergence of large-scale vision-language models, there has been a surge in text generation models for medical imaging tasks, particularly for radiology report generation. However, this growing diversity and expressiveness of generated reports

poses new challenges for evaluation. Traditional natural language generation metrics—such as BLEU or ROUGE—primarily rely on surface-level word overlap, which often fails to capture the semantic or clinical correctness of generated outputs. This limitation becomes especially pronounced in the medical domain, where subtle differences in terminology can carry significant clinical implications.

This thesis focuses on improving performance in three core areas of medical vision-language research: **Visual Question Answering (VQA)**, **Radiology Report Generation (RRG)**, and **Radiology Report Evaluation Metrics**. Despite the growing adoption of deep learning methods in these domains, existing models often face significant limitations—particularly in capturing complex medical semantics, structured reasoning, and aligning generated outputs with expert-level evaluation standards.

In medical visual question answering (VQA), questions can generally be categorized into two types: **close-ended** and **open-ended**. Close-ended questions typically require short, categorical responses such as "yes", "no", or specific disease names (e.g., "Is there evidence of pneumonia?"). In contrast, open-ended questions demand more descriptive answers, often involving detailed reasoning or summarization based on complex imaging findings (e.g., "Describe the abnormalities seen in the left lung."). For **medical VQA**, most existing approaches follow a classification-based paradigm, where each answer is treated as a discrete class label. While effective for close-ended questions, this approach is inflexible for open-ended tasks and fails to capture rich semantic dependencies between the image, question, and answer. Generative models address this by producing free-form answers token-by-token, but they tend to suffer from hallucination and lack grounding in clinical concepts. To overcome these limitations, we propose **Q2ATransformer**, a semi-open framework that unifies classification and generation paradigms. It introduces trainable answer candidate embeddings that interact with multi-modal representations via a Transformer decoder. This enables the model to combine the precision of classification with the semantic flexibility of generation, improving performance on both close- and open-ended medical questions.

**Radiology report generation (RRG)** aims to generate comprehensive diagnostic reports from medical images such as chest X-rays or CT scans. Most existing methods adopt an encoder-decoder architecture, where a visual encoder extracts image features and a language decoder (often a Transformer or LLM) autoregressively generates the textual report. For **RRG**, current models typically rely on cross-entropy loss and autoregressive decoding, which treat all tokens equally and penalize semantically valid variations. This leads to brittle training dynamics and poor factual correctness. Moreover, most training pipelines lack feedback mechanisms for identifying and correcting specific semantic errors. To address this, we introduce **SAT-RRG (Self-Adaptive Training for RRG)**, a novel learning framework that enables token-level semantic supervision without requiring external annotations. By leveraging the intrinsic understanding of a large language model, SAT-RRG automatically identifies error-prone tokens within generated reports and applies two complementary loss functions: *Correct Token Augment Loss (CTAL)* for stabilizing accurate predictions, and *Error Token Adaptive Penalty Loss (ETAPL)* for correcting semantically incorrect tokens. This strategy enhances both fluency and clinical accuracy of generated reports.

Motivated by the strong generalization capabilities of GPT-4V, we also conduct a systematic evaluation of its zero-shot and few-shot performance across medical VQA, RRG, and VG tasks. GPT-4V has demonstrated impressive multimodal reasoning abilities in general-domain tasks such as image captioning, visual question answering, and document understanding. However, medical imaging poses unique challenges that differ significantly from general vision-language benchmarks. These include the need for fine-grained anatomical understanding, recognition of subtle pathological patterns, and alignment with domain-specific clinical knowledge. Therefore, it remains an open question whether GPT-4V’s impressive general capabilities can transfer to the highly specialized domain of medical vision-language tasks. Our findings reveal that while GPT-4V produces clinically plausible and diverse outputs, existing automatic evaluation metrics—such as BLEU, ROUGE, and CIDEr—struggle to accurately assess quality due to their reliance on surface-level n-gram

overlap. These metrics often fail to correlate with human expert ratings, especially in scenarios where multiple semantically correct but lexically diverse answers exist.

To address these evaluation challenges, we propose **MRScore**, a reward-based metric powered by LLMs that aligns more closely with radiologists' judgment. MRScore enables customizable evaluation criteria (e.g., factual accuracy, completeness, clinical relevance) and operates on the semantic level to better assess correctness in generated reports. However, a single score alone lacks diagnostic utility, prompting us to further introduce **ReFINE**—an interpretable evaluation framework that decomposes the overall score into multiple fine-grained sub-scores. These sub-scores provide feedback on specific weaknesses, allowing model developers to identify and address deficiencies under each criterion.

In summary, this thesis investigates both the capability of pretrained generalist VLMs in the medical domain and the limitations of current evaluation frameworks. By proposing novel models and evaluation metrics, we aim to advance the reliability, interpretability, and clinical applicability of vision-language systems in medical image analysis.

## 1.2 Challenges and Motivations

### 1.2.1 Challenges

#### **Challenge 1: Medical Image Visual Question Answering (VQA)**

Medical Visual Question Answering (VQA) is a multimodal task that requires reasoning over image content to answer natural language questions. It typically involves four stages: extracting image features, extracting question features, integrating the two modalities, and generating or selecting an appropriate answer. While general-domain VQA systems have made significant progress [43, 111], medical VQA introduces additional complexities due to the need for domain-specific knowledge and fine-grained visual understanding.

In medical VQA, questions may be close-ended (e.g., “Is there a pleural effusion?”) or open-ended (e.g., “Describe the major findings”), and current methods often fail to handle both types effectively. Most existing approaches [75, 26, 34, 29, 28] use a classification-based framework, treating each possible answer as a discrete label. While suitable for close-ended questions, this rigid structure struggles with open-ended ones due to the diversity and length of expected responses.

On the other hand, generative methods [1, 49] treat VQA as a sequence generation task and can provide more expressive and descriptive answers. However, they suffer from an expansive output space and often produce hallucinated or clinically irrelevant content, resulting in lower factual accuracy.

Some hybrid models attempt to address this by treating close-ended questions with classification and open-ended questions with generation [89], but such approaches lack architectural unity and fail to capture shared reasoning structures across question types.

### **Challenge 2: Radiology Report Generation (RRG)**

Automated radiology report generation (RRG) has become a significant research focus due to its potential to reduce radiologists’ workload and improve diagnostic standardization. With the rise of deep learning, particularly encoder-decoder architectures and large language models (LLMs) [98, 17, 16, 103, 39, 10], substantial progress has been made in generating clinically relevant reports from radiological images.

Despite these advances, current RRG systems predominantly rely on cross-entropy (CE) loss for training. In this framework, each token in the report is predicted autoregressively, and any deviation from the reference report is penalized—even when the generated token is semantically equivalent. This rigid supervision can discourage diverse yet clinically correct expressions and introduces a bias toward surface-level matching.

Moreover, most existing models lack a mechanism to detect and correct factual or semantic errors during training. They treat all token errors equally, regardless of their clinical significance. This can lead to

fluent but factually incorrect outputs, which pose serious risks in medical applications.

In the broader natural language generation (NLG) field, reinforcement learning [129], contrastive training [62], and correction-based fine-tuning have been proposed to improve semantic alignment. However, these methods often require manually designed reward functions or expert-labeled training signals, making them difficult to scale in clinical domains where annotations are expensive and limited.

### **Challenge 3: Evaluating Large VLMs for Zero-shot and Few-shot Medical Image Understanding**

Recent advances in large language models (LLMs) [96, 77, 2] and their multimodal extensions (LMMs) [119, 53, 3] have opened up new possibilities for zero-shot and few-shot image understanding. Models such as GPT-4V have demonstrated impressive generalization to visual tasks across domains, including captioning, question answering, and visual reasoning, all without extensive task-specific fine-tuning.

However, their performance in the medical domain remains under-explored and, based on initial findings, limited in key areas. Unlike natural images, medical images are grayscale, high-resolution, and domain-specific, often requiring subtle pattern recognition and expert-level clinical interpretation. This raises concerns about the ability of generalist VLMs to perform complex medical reasoning tasks out of the box.

In particular, the zero-shot and few-shot capabilities of GPT-4V in three representative medical tasks—radiology report generation, medical visual question answering (VQA), and visual grounding—remain unclear. While early results are promising in terms of linguistic fluency and general reasoning, the models often struggle with clinical correctness, spatial localization, and consistency with radiologist judgment.

#### **Challenge 4: Limitations of Traditional Radiology Report Evaluation Metrics**

Evaluating the outputs of automated radiology report generation (RRG) systems is a non-trivial task that requires accurate judgment of both semantic meaning and clinical relevance. However, widely adopted natural language generation (NLG) metrics—such as BLEU [82], METEOR, and ROUGE—primarily rely on surface-level n-gram overlap with ground-truth references. While effective in general NLP tasks, these metrics fail to capture the flexibility and clinical nuance inherent to radiology reports.

First, such string-matching metrics penalize valid paraphrasing. For instance, "cardiomegaly" and "enlarged heart" convey the same finding, yet receive low similarity scores due to lexical variation. This leads to misleading evaluations when semantically correct reports use different phrasing from the ground truth.

Second, traditional metrics are agnostic to clinical correctness. They cannot differentiate between stylistic variation and factual errors, nor do they prioritize medically important content. To compensate, domain-specific metrics like CheXbert [92] and RadGraph [41] have been proposed, which measure the correctness of extracted clinical entities and relations. However, these are still limited by fixed schemas and do not assess linguistic fluency or contextual coherence.

Third, and most critically, existing metrics show poor correlation with human expert ratings [56]. This discrepancy undermines their reliability for benchmarking, especially in high-stakes clinical applications. Composite metrics like RadCliQ [120] attempt to improve alignment by learning from human-labeled quality scores using regression over multiple metrics. However, these methods depend heavily on costly expert annotations and are biased toward specific label distributions.

#### **Challenge 5: Lack of Fine-Grained and Practical Evaluation in Existing Radiology Report Metrics**

Automated radiology report generation (R2Gen) has made notable strides in recent years [105, 57], yet evaluating the quality of generated reports

remains a persistent bottleneck. The complexity of this task lies in its need to capture subtle clinical semantics, correct diagnostic details, and structured report conventions—all of which are difficult to assess using conventional natural language generation (NLG) metrics.

Metrics such as BLEU [82], METEOR [88], and ROUGE are widely used in general NLG, but they rely heavily on n-gram overlap and are blind to paraphrasing and structural variation—features that are common and often clinically acceptable in medical reporting. More advanced methods such as BERTScore [124] mitigate this through contextual embeddings, but they still fall short in evaluating factual accuracy or medical validity.

Clinically grounded metrics like CheXbert [92] and RadGraph F1 [41] improve upon this by checking for the presence and correctness of extracted medical entities and relations. However, they are limited to specific schemas and offer only coarse-grained views of report quality, ignoring other important dimensions such as fluency, completeness, or severity attribution.

Even composite metrics like RadCliQ [121], which combine multiple scores via regression models trained on expert-labeled data, suffer from two major limitations: (1) reliance on scarce, expensive human annotations; and (2) lack of flexibility in adapting to new evaluation criteria or clinical domains.

LLM-based evaluation methods (e.g., GPT-4 [77]) present a promising alternative. However, most of them require online API access, raising concerns about privacy, inference cost, and reproducibility—especially in sensitive clinical settings.

## 1.2.2 Motivations

### **Motivation 1: Towards Unified and Accurate Medical VQA**

To address the limitations of existing classification-based and generation-based VQA approaches, we are motivated to develop a unified, semi-open framework that combines the strengths of both paradigms. Our

proposed model, Q2ATransformer, retains the structured answer space of classification models while introducing semantic reasoning mechanisms from generative models.

Specifically, we introduce trainable answer candidate embeddings that interact with the fused image-question representation via a Transformer decoder. This structure enables the model to incorporate semantic context and produce more informed, accurate predictions. Additionally, we propose a Cross-Modality Attention Network (CMAN) to improve feature fusion. Unlike traditional element-wise fusion methods, CMAN uses self-attention over concatenated features to better capture inter-modal relationships.

Our model demonstrates significant performance gains, especially for open-ended questions—a known bottleneck in prior work. On two benchmark datasets (VQA-RAD and PathVQA), our method achieves new state-of-the-art results, with notable absolute gains of 16% and 41% respectively on open-ended question tasks.

These outcomes reinforce our motivation to design VQA systems that are not only more flexible and accurate but also semantically aware and clinically reliable.

### **Motivation 2: Improving Radiology Report Generation with Self-Adaptive Feedback**

Motivated by the need to move beyond rigid CE loss and external reward engineering, we propose a self-supervised, semantically aware training framework for RRG. Our goal is to enable models to autonomously identify and correct their own semantic errors during training by leveraging the intrinsic understanding of a large language model backbone.

At the heart of our framework is an internal error detection module, which compares the predicted and reference reports to locate tokens that are semantically inconsistent. This feedback does not require external annotations, allowing the model to refine itself in a scalable and efficient manner.

Based on this internal signal, we introduce two novel loss functions:

- **Correct Token Augment Loss (CTAL):** Encourages stability by minimizing updates for tokens that are predicted correctly, helping the model retain good predictions.
- **Error Token Adaptive Penalty Loss (ETAPL):** Applies adaptive penalties to incorrect tokens based on their confidence, increasing corrective pressure only where needed.

This design allows the model to balance the preservation of reliable content with the targeted correction of factual mistakes. Importantly, it introduces no architectural or inference-time overhead, making it practical for real-world deployment.

We demonstrate that this framework achieves state-of-the-art results on two major datasets—MIMIC-CXR and IU-Xray—while offering better factual alignment, improved clinical validity, and robust generalization without the need for external rewards or human-labeled correction data.

### **Motivation 3: Benchmarking GPT-4V in Medical Vision-Language Tasks**

Motivated by the rapid adoption of GPT-4V and similar large vision-language models, we aim to systematically evaluate their capabilities in high-impact medical imaging tasks. While prior studies have hinted at their potential, rigorous and comprehensive evaluations—especially from a clinical perspective—are lacking.

To address this, we design experiments across three core vision-language tasks: radiology report generation, medical VQA, and visual grounding. Each task probes different aspects of the model’s capability, from free-text summarization to precise region-level understanding. Our evaluation includes:

- **Quantitative analysis** against domain-specific baselines using standard metrics;
- **Expert-based human evaluation** to assess fluency, accuracy, and clinical value;

- **Case studies** to illustrate strengths, limitations, and typical failure modes.

We also explore different prompting strategies tailored for each task, including both zero-shot and few-shot settings. This allows us to study how prompt construction impacts model behavior and whether prompting alone can activate medical reasoning capabilities in generalist models.

Our findings suggest that:

- GPT-4V produces clinically relevant and linguistically fluent reports, sometimes outperforming traditional metrics in human evaluations.
- It answers VQA questions with reasonable detail and contextual awareness, but suffers from formatting discrepancies that hinder metric-based performance.
- It performs poorly on visual grounding tasks, highlighting a major gap in its spatial understanding of medical images.
- Human evaluations often score GPT-4V's outputs higher than automated metrics do, revealing a persistent discrepancy and motivating the development of better evaluation tools.

These insights motivate a deeper investigation into the domain adaptation, evaluation criteria, and practical readiness of large generalist VLMs in the medical domain.

#### **Motivation 4: Developing Human-Aligned and Scalable Evaluation Metrics**

Motivated by the shortcomings of traditional and composite evaluation metrics, we propose a new approach to radiology report evaluation that leverages large language models (LLMs) for scalable, human-aligned quality assessment. Our goal is to create a metric that better reflects radiologist judgment, operates at a semantic level, and can be adapted to different clinical needs—all without requiring extensive expert-labeled data.

To this end, we introduce **MRScore**, a novel reward-based evaluation metric tailored specifically for radiology report generation. Built in collaboration with radiologists, MRScore defines a structured framework of seven evaluation dimensions, such as factual accuracy, completeness, and clinical relevance. Rather than relying on expensive human annotations, we use GPT-4 [77] to generate synthetic report variants labeled as low, medium, or high quality across each dimension. From 1,000 reference reports, this process yields 3,000 quality-controlled samples.

These samples are then organized into <accepted, rejected> preference pairs and used to train a reward model on top of Mistral-7B-Instruct [42]. The result is a lightweight, fine-tuned evaluation model, MRScore, that produces scalar reward outputs aligned with radiologist-like assessments.

**Our approach offers several advantages:**

- It eliminates the need for large-scale manual annotations by simulating human evaluation through GPT-4 prompting with a few-shot in-context setup.
- It supports scalable evaluation via a trained reward model, enabling efficient comparison across large numbers of reports.
- It improves alignment with human ratings, as shown through correlation analysis on GPT-4V-generated reports.

By designing an interpretable, LLM-powered evaluation framework that reflects domain-specific quality criteria, we aim to close the gap between automatic metrics and real clinical expectations.

**Motivation 5: Toward Fine-Grained, Interpretable, and Deployable Evaluation with ReFINE**

To address the limitations of current radiology report evaluation metrics, we are motivated to develop a new framework that is: (1) fine-grained and interpretable; (2) aligned with expert judgment; (3) scalable and trainable without manual annotation; and (4) efficient enough for practical deployment.

We introduce **ReFINE**, a novel evaluation metric specifically tailored for R2Gen. Unlike prior approaches, ReFINE provides both an overall quality score and interpretable sub-scores across user-defined clinical criteria. This enables detailed insight into model performance dimensions—such as whether errors stem from missing findings, wrong anatomical location, or misattributed severity.

Our method is built around a margin-based reward enforcement loss, which allows a reward model to learn from pairwise quality comparisons while maintaining training stability. We generate training data using GPT-4 to simulate expert judgment across seven evaluation criteria. For each of 1,000 reference reports, we prompt GPT-4 to produce three quality levels (low, medium, high), then construct  $\langle \text{accepted}, \text{rejected} \rangle$  pairs for training a reward model based on Mistral-7B-Instruct [42].

**Compared with prior methods, ReFINE offers several distinct advantages:**

- *Customizability*: Unlike static metrics, ReFINE’s sub-scores reflect predefined clinical dimensions and can be adjusted to suit different use cases.
- *Interpretability*: Instead of a single score, ReFINE provides a vector of interpretable feedback, enabling targeted error diagnosis and model refinement.
- *Efficiency and privacy*: ReFINE is trained and deployed offline on a single GPU (A6000), avoiding dependency on proprietary APIs or cloud access.
- *Superior alignment*: ReFINE achieves a higher Kendall’s Tau correlation (0.75) with expert scores than GREEN (0.64), while being faster and lighter to train and run.

We validate ReFINE across three diverse expert-annotated datasets, **ReXVal**, **RaTE-Eval**, and our newly collected **Rad-100**, and consistently observe better correlation with radiologist ratings than existing alternatives. Through this, ReFINE establishes a new paradigm for reliable, fine-grained, and practical evaluation in radiology report generation.

### 1.3 Thesis Contribution and Outline

This thesis focuses on three core pillars of medical vision-language modeling: **Medical Visual Question Answering (VQA)**, **Radiology Report Generation (RRG)**, and **Evaluation Metrics for Radiology Reports**. First, we advance the modeling of medical VQA by proposing a unified framework that balances the strengths of both classification and generation paradigms. Second, we introduce a self-adaptive training mechanism for radiology report generation that enables fine-grained, semantically aware token-level learning. Third, we develop novel evaluation metrics tailored to clinical report quality, including both scalar (MRScore) and fine-grained (ReFINE) scoring approaches, to better reflect radiologist judgment.

Collectively, this research aims to bridge the gap between general-purpose multimodal models and the specialized needs of medical imaging tasks, offering new tools for both generation and evaluation in clinical AI systems.

**The main contributions of this thesis are as follows:**

- **Q2ATransformer:** This thesis proposes **Q2ATransformer**, a semi-open solution that combines classification and generation paradigms through answer-aware semantic reasoning. The model introduces trainable answer embeddings and a Transformer decoder for flexible yet constrained answer space exploration, achieving state-of-the-art results on VQA-RAD and PathVQA benchmarks.
- **Self-Adaptive Training for RRG:** To overcome the rigidity of cross-entropy loss, a self-supervised training framework is developed that integrates token-level semantic feedback using the model’s own capabilities. Two novel loss functions, Correct Token Augment Loss (CTAL) and Error Token Adaptive Penalty Loss (ETAPL), are introduced to distinguish between correct and semantically erroneous tokens, improving clinical report fidelity.
- **Evaluation of Generalist LMMs in Medical Tasks:** GPT-4V is systematically evaluated on three medical vision-language tasks:

report generation, VQA, and visual grounding. The analysis includes metric-based and human expert evaluations, exposing both the strengths (e.g., contextual fluency) and weaknesses (e.g., visual grounding limitations) of generalist models in clinical use.

- **MRScore: A Human-Aligned Evaluation Metric:** This work introduces MRScore, a reward model trained on synthetic report pairs generated by GPT-4 with human-like scoring. The model captures radiologist evaluation logic and demonstrates superior correlation with expert ratings compared to existing metrics.
- **ReFINE: Fine-Grained Report Evaluation Metric:** ReFINE is proposed as a criterion-aware reward model using a margin-based loss that enables interpretable sub-score decomposition across multiple evaluation dimensions. ReFINE outperforms prior methods (e.g., RadCliQ, GREEN) in both scoring accuracy and explainability across three expert-labeled datasets.

#### Thesis Outline:

- **Chapter 1 – Introduction:** Presents the motivation, background, and high-level challenges in medical vision-language research.
- **Chapter 2 – Literature Review:** Surveys relevant works in medical VQA, radiology report generation, evaluation metrics, and general-purpose multimodal models.
- **Chapter 3 – Q2ATransformer:** a semi-open framework for medical VQA that unifies the strengths of classification- and generation-based models, enabling more accurate and semantically expressive question answering..
- **Chapter 4 – SAT-RRG:** a self-adaptive training framework for radiology report generation that leverages in-model semantic feedback to improve token-level supervision without requiring human annotations.
- **Chapter 5 – Evaluation of GPT-4V on Medical Tasks:** conduct a systematic evaluation of GPT-4V on three representative medical

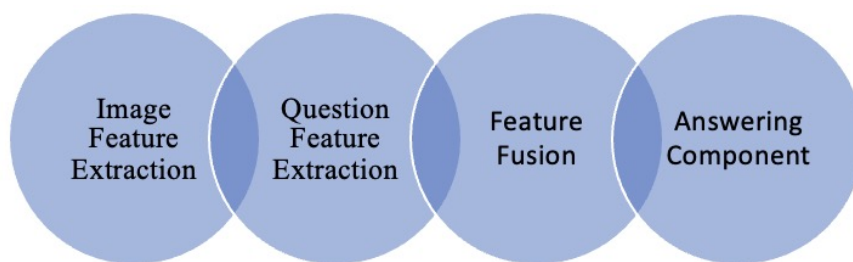
vision-language tasks: radiology report generation, medical VQA, and visual grounding. Our analysis highlights both its strengths and its current limitations in the medical domain.

- **Chapter 6 – MRScore:** a human-aligned, LLM-based evaluation metric trained with pseudo-labeled reward signals to improve alignment with expert assessments in RRG.
- **Chapter 7 – ReFINE:** a fine-grained evaluation framework that decomposes report quality into sub-criteria scores, enabling interpretable, flexible, and criterion-specific evaluation for radiology reports.
- **Chapter 8 – Conclusion and Future Work:** Summarizes the key findings and outlines possible extensions, including multimodal grounding, domain adaptation, and real-world deployment.

## Chapter 2

# Literature Review

In this chapter, we begin by introducing the background of three core applications in medical vision-language modeling: visual question answering (VQA), radiology report generation (RRG), and visual grounding (VG). We then review the recent advancements and challenges in each task, highlighting the need for domain-specific understanding, clinical reasoning, and multimodal alignment. Following this, we focus on evaluation metrics for RRG, analyzing commonly used automatic metrics such as BLEU and ROUGE, and discussing their limitations in capturing semantic correctness and clinical relevance. This chapter provides a comprehensive overview of medical vision-language tasks and RRG evaluation metrics, laying the groundwork for the methodologies proposed in this thesis.



**Figure 2.1:** The general pipeline of a medical Visual Question Answering (VQA) model. It includes four key components: image feature extraction, question feature extraction, feature fusion, and the final answer generation.

## 2.1 Visual Question Answering

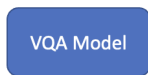
Visual Question Answering (VQA) is a multimodal task at the intersection of computer vision and natural language processing, which aims to automatically answer natural language questions based on image content. A typical VQA system consists of four core modules shown in Figure 2.1: visual feature extraction, question encoding, multimodal fusion, and answer prediction. The task has seen significant advancements in recent years with the development of deep learning techniques [43, 111].

Medical VQA extends the general VQA framework to the clinical domain by requiring models to answer questions grounded in medical images. Unlike general VQA, medical VQA demands a deeper understanding of domain-specific visual patterns and clinical knowledge. It offers a promising way to assist clinical decision-making by enabling a natural and interpretable form of interaction. Medical VQA tasks typically involve both close-ended (e.g., Yes/No) and open-ended questions, which differ significantly in complexity and answer space, there are two examples of close-ended and open-ended questions in Figure 2.2.

### ➤ Yes/No(Closed) Question Type

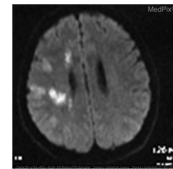


“Are the lungs normal appearing?”



“No”

### ➤ What/How(Open-end) Question Type



“What type of imaging is this?”



“MRI Diffusion Weighted”

**Figure 2.2:** Two common types of questions in medical Visual Question Answering (VQA). Left: Yes/No (Close-ended) question, where the model answers binary clinical queries. Right: What/How (Open-ended) question, where the model generates descriptive clinical answers.

Early works in medical VQA primarily adopt closed-type approaches,

in which the task is formulated as a multi-class classification problem [75, 26, 34, 29, 28]. As illustrated in Figure 1(a), these models directly predict answers from a predefined answer set based on the fused image-question representations. This formulation simplifies the task and enables strong performance on close-ended questions, especially binary ones. However, the limitation lies in its inability to handle open-ended questions effectively due to the rigid answer space and lack of semantic modeling of answers.

To address this, open-type approaches have been proposed [1, 49], which reformulate medical VQA as a sequence generation task (Figure 1(b)). These models generate answers token-by-token using encoder-decoder architectures, allowing them to handle more expressive and semantically rich answers. While promising, these approaches often suffer from low accuracy and incoherence due to the vast search space of possible outputs and limited annotated training data.

Hybrid models that attempt to combine the strengths of both classification and generation frameworks have also emerged. For instance, CGMVQA [89] treats close-ended and open-ended questions with separate pipelines, applying classification for the former and generation for the latter. However, this hard separation prevents the model from sharing semantic knowledge across different types of questions, limiting generalizability.

To bridge this gap, we introduce a semi-open type framework, Q2A-Transformer (Figure 1(c)). It maintains the classification backbone for answer prediction but augments it with semantic modeling of answer embeddings. Specifically, learnable answer embeddings are treated as queries in a transformer decoder, which attends over fused image-question features to capture fine-grained answer relevance. This design enables bidirectional interaction between image-question and answer semantics, a characteristic traditionally found only in generation-based methods. Compared with existing approaches, our method improves answer expressiveness without incurring the instability of free-form generation, offering a better trade-off between interpretability, accuracy, and flexibility in medical VQA.

## 2.2 Radiology Report Generation

Radiology report generation (RRG) aims to automatically produce diagnostic narratives from medical images, bridging the gap between visual understanding and clinical language. Early approaches typically adopt encoder-decoder frameworks trained with word-level cross-entropy (CE) loss [17, 66]. While effective to some extent, this token-level supervision treats all tokens equally and penalizes semantically valid paraphrases, limiting linguistic diversity and clinical flexibility. To overcome these shortcomings, recent research has introduced domain-specific knowledge and novel architectural designs. For instance, METransformer [104] employs learnable “expert tokens” to enhance cross-modal attention alignment, EKAGen [10] integrates instance-level expert knowledge to better supervise attention distributions, and KiUT [39] applies a U-shaped multi-level modality connection to fuse clinical semantics across encoder and decoder stages. Despite these innovations, most of these models still apply coarse-grained supervision and fail to distinguish between clinically accurate and erroneous content at a fine-grained level during training.

With the advent of large language models (LLMs), RRG has entered a new stage. Pretrained autoregressive models like LLaMA [96] and Qwen [114] have demonstrated strong text generation abilities, and recent adaptations to the radiology domain have yielded promising results. For example, R2GenGPT [105] introduces a linear visual mapper to adapt image features to the LLM input space, while Bootstrapping [64] incorporates in-domain prompts and coarse-to-fine decoding strategies to better control report generation. However, these LLM-based methods typically inherit the limitations of their predecessors, lacking mechanisms to provide token-level feedback or to distinguish between correct and incorrect predictions during training.

In natural language generation (NLG), fine-grained supervision has been explored through reinforcement learning (RL) with human preferences [129], contrastive learning [62], and post-hoc correction-based

methods. Yet, such strategies often require extensive labeled data or reward design, making them less scalable, particularly in medical contexts. Within medical NLP, token-level semantic feedback remains underexplored due to high annotation costs and the lack of effective automated mechanisms.

To address these challenges, recent efforts, including our own, propose leveraging frozen LLMs for in-context prompting to detect semantically inconsistent tokens automatically. This removes the need for manual annotation while enabling scalable, fine-grained supervision. Coupled with tailored loss functions—such as error-aware penalization and confidence-based reinforcement—this direction offers a new path forward for improving both the accuracy and interpretability of radiology report generation.

## 2.3 Visual Grounding

We introduce the task of **visual grounding** here as it serves as a core component in our later evaluation of GPT-4V’s capabilities (Chapter 5), particularly in assessing its spatial understanding of medical images. Visual grounding (VG) is a vision-language task that aims to associate natural language descriptions with specific regions in an image by outputting the coordinates of a bounding box corresponding to the described object or concept [47]. In the medical domain, this task translates into localizing radiologically significant findings—such as lesions, consolidations, or implants—based on descriptive statements that often reflect clinical symptoms or diagnostic interpretations. For example, a model may be required to identify and demarcate a region exhibiting “right lower lobe consolidation” or “cardiomegaly” from a chest X-ray image. Compared to radiology report generation and visual question answering, visual grounding is unique in that its output is not textual but spatial, posing distinct challenges in aligning semantic descriptions with visual evidence.

Most prior research in visual grounding has primarily focused on natural images, leveraging large-scale datasets such as RefCOCO [48],

and achieving remarkable progress through transformer-based models like MDETR [47]. These models benefit from abundant annotated data and a clear correspondence between textual phrases and visual objects. In contrast, medical visual grounding remains an underexplored frontier, largely due to the scarcity of annotated datasets that offer paired medical descriptions and precise region-level annotations. As a result, models trained in natural image domains often fail to generalize to the specialized semantics and fine-grained anatomical details present in medical imagery.

Recently, the release of the MS-CXR dataset has provided a substantial boost to this research area, offering paired chest X-rays and annotated phrases with corresponding bounding boxes. This dataset has enabled several pioneering studies in medical visual grounding. For example, Huang et al.[38] propose a grounding framework that leverages both global and local image-text alignments, while Sun et al.[94, 95] explore contrastive and retrieval-based formulations for localizing medical findings. These studies demonstrate the feasibility of applying visual grounding techniques in clinical settings, but also highlight the need for models capable of understanding subtle anatomical cues and domain-specific terminology.

In parallel, the integration of large multimodal models (LMMs), particularly those extending large language models (LLMs) with visual encoders, has shown great promise in general-domain VG. Models like Kosmos-2 [86] and BuboGPT [127] have demonstrated the capability to perform grounding-like tasks by directly generating bounding box coordinates from textual prompts and images, effectively bridging semantic understanding and visual localization. These approaches suggest that LLMs, when properly adapted, can internalize spatial reasoning without the need for dense supervision.

Despite these advancements, the application of such models in the medical domain remains limited. Given the nuanced and specialized nature of clinical language and visual patterns, it is unclear to what extent general-purpose LMMs can generalize to medical visual grounding. This motivates the need for systematic evaluation of foundation models such

as GPT-4V, which integrates powerful vision-language reasoning capabilities. Understanding whether GPT-4V can interpret medical images and produce clinically meaningful spatial outputs—without task-specific fine-tuning—would provide valuable insight into its zero-shot grounding ability and potential as a diagnostic assistant.

## 2.4 Radiology Report Evaluation Metrics

Evaluating radiology report generation models presents unique challenges due to the need to assess both linguistic fluency and clinical accuracy. Existing evaluation metrics can broadly be categorized into language-based and clinical-based metrics. The most commonly used traditional radiology report evaluation metrics were listed in Table 2.1 and the LLM-based metrics were listed in Table 2.2

Language metrics assess the textual similarity between the generated report and the ground truth reference, relying primarily on lexical overlap or embedding-based similarity. Classic metrics in this category include BLEU [82], which measures n-gram precision; ROUGE [61], which focuses on recall of overlapping sequences; and METEOR [6], which integrates synonymy and stemming. More recently, BERTScore [123] leverages contextualized word embeddings from pre-trained transformers to compute semantic similarity. While these metrics are well-established in general-purpose natural language generation, they often fall short in radiology-specific contexts. Medical reports frequently contain semantically equivalent but lexically diverse expressions (e.g., “no focal consolidation” vs. “lungs are clear”), which such metrics may erroneously penalize. As a result, high scores on these metrics do not necessarily correspond to clinically meaningful accuracy.

To address this gap, clinical metrics have been introduced to evaluate the factual correctness of generated reports. CheXpert [40] and its extension CheXbert utilize rule-based or model-based labelers to extract the presence, absence, or uncertainty of 14 predefined pathologies. These labels are then compared to reference annotations for evaluation. Another prominent clinical metric, RadGraph [41], parses radiology reports

| Name               | Type               | Focus/Usage                 | How to Compute / Example   |
|--------------------|--------------------|-----------------------------|--|
| BLEU[81]           | Linguistic         | n-gram overlap, fluency     | Measures the precision of n-grams (up to 4) in the candidate text against reference texts. Calculated using the geometric mean of n-gram precisions and a brevity penalty.               |
| ROUGE-L[61]        | Linguistic         | Recall, n-gram coverage     | Measures the longest common subsequence (LCS) between candidate and reference texts, capturing sentence-level structure similarity. Computes F1 score based on LCS precision and recall. |
| METEOR[6]          | Linguistic         | Semantic, synonym, stem     | Considers word matches, stemming, synonyms; computes harmonic mean of precision and recall.  |
| CIDEr[100]         | Linguistic         | Consensus/image description | TF-IDF weighted n-gram similarity, designed for image captions; higher for more "human-like" descriptions.   |
| BERTScore[123]     | Linguistic         | Semantic similarity         | Uses BERT embeddings to compute similarity between candidate and reference texts at a token level. Measures precision, recall, and F1 score.   |
| CheXpert Score[40] | Clinical           | Finding-level agreement     | Extracts clinical labels (e.g., pneumonia, cardiomegaly) from both reports, calculates F1-score for each label.  |
| RadGraph Score[41] | Clinical           | Entity/Relation accuracy    | Uses BERT embeddings to compute similarity between candidate and reference texts at a token level. Measures precision, recall, and F1 score.   |
| RadCliQ[121]       | Composite Clinical | Overall clinical quality    | Aggregates various automated metrics, trained to correlate with radiologist judgments.   |

**Table 2.1:** Traditional evaluation metrics for radiology report generation.

| Name             | Type      | Focus/Usage   | How to Compute / Example  |
|------------------|-----------|---|---|
| GREEN[78]        | LLM-based | Factual, clinical, linguistic consistency                           | Uses an LLM (e.g., GPT-4) to rate the generated report on factual agreement, clinical accuracy, and linguistic quality.   |
| RaTEScore[126]   | LLM-based | Entity-level factual accuracy with clinical weighting               | Evaluates radiology reports by extracting medical entities using a specialized NER model, then computing embedding-based similarity between matched entities, with clinical importance used to weight the final score. This allows it to assess factual accuracy while being robust to synonyms and wording variations.   |
| G-Rad Score[13]  | LLM-based | Clinically categorized error identification                         | Evaluates generated reports by prompting GPT-4T to count six types of errors (e.g., false prediction, omission, location error, severity error, etc.), distinguishing between clinically significant and insignificant cases.   |
| RadFact[7]       | LLM-based | Logical and spatial factual consistency (with or without grounding) | Evaluates radiology reports by prompting an LLM (Llama3-70B-Instruct) to verify whether the ground-truth report logically entails each generated sentence, computing logical precision and recall to assess factuality and completeness. When grounding annotations are available, it further assesses spatial grounding accuracy, without relying on predefined error types or specialized radiology models. |
| CheXprompt[12]   | LLM-based | Clinically categorized error identification                         | Prompts GPT-4T to count six types of errors (e.g., false prediction, omission, location error, severity error, etc.), distinguishing between clinically significant and insignificant cases.  |
| FineRadScore[37] | LLM-based | Line-level correction and clinical severity assessment              | Evaluates generated reports by prompting an LLM to identify line-by-line corrections needed to match a reference report, labeling each correction with an edit type (insertion, deletion, substitution), a clinical severity level, and one or more of six ReXVal error categories. The model's outputs are then compared to radiologist annotations to assess alignment.                                     |

**Table 2.2:** LLM-based evaluation metrics for radiology report generation.

into structured graphs of clinical entities and their relations, allowing for a relational correctness assessment. Despite their utility, such extraction-based approaches are constrained by limited ontology coverage and rigid matching rules, rendering them less effective for nuanced or ambiguous findings. Additionally, they may struggle with novel terminology or rare conditions not covered in the predefined label sets.

To mitigate the limitations of purely linguistic or purely clinical metrics, several hybrid approaches have emerged. For example, RadCliQ [121] combines multiple extraction-based evaluations into a weighted composite score, and RadEval [11] integrates both rule-based and neural similarity scoring. Although these methods improve alignment with human judgment to some extent, they still exhibit limited adaptability to diverse report styles and do not provide fine-grained diagnostic feedback.

## Chapter 3

# Q2ATransformer: Improving Medical VQA via an Answer Querying Decoder

In this chapter, we conduct an in-depth exploration of enhancing the effectiveness of Medical Visual Question Answering (VQA) systems. These systems play a complementary role in interpreting clinically significant information embedded within medical images. Medical image-related questions generally fall into two primary types: closed-end (e.g., Yes/No questions) and open-end queries. Most existing methods for medical VQA treat this task as a classification problem, although some recent efforts attempt generative approaches or hybrid frameworks—utilizing classification for closed-end questions and generation for open-ended ones. While classification-based models are straightforward to implement, they tend to struggle with long, descriptive open-ended questions. Conversely, generation-based methods, despite offering more flexible outputs, often produce irrelevant or non-existent answers, thus compromising accuracy. To address these limitations, we introduce a novel Transformer-based architecture named **Q2ATransformer** that unifies the benefits of both paradigms. Our approach treats both question types within a single framework by incorporating an additional Transformer decoder equipped with a set of learnable candidate answer embeddings. These embeddings query the presence of potential answer classes for a given image-question pair. Through the attention mechanisms in the Transformer, the candidate embeddings dynamically interact with the

fused representation of the image and question to facilitate prediction. Although the overall framework follows a classification structure, it mimics the semantic interaction of generative models with answer content. Simultaneously, by constraining the answer space through classification, our method reduces the complexity of the prediction task. Experimental results show that our approach achieves new state-of-the-art results on two benchmark datasets in medical VQA. In particular, for open-ended questions, our model reaches 79.19% accuracy on VQA-RAD and 54.85% on PathVQA, representing absolute improvements of 16.09% and 41.45%, respectively.

### 3.1 Introduction

Visual Question Answering (VQA) is widely recognized as a complex AI task that involves providing accurate responses to questions based on the visual content of images. This task necessitates the joint use of computer vision and natural language processing techniques, typically comprising four essential stages: image feature extraction, question feature extraction, multimodal feature fusion, and answer prediction. In recent years, notable advancements have been made in this field [43, 111]. Medical VQA naturally extends traditional VQA into the domain of healthcare by focusing on medical images accompanied by clinically relevant questions. Through a question-answering interface, it offers an intuitive and interactive way to support clinical decision-making. The posed questions in medical VQA can be categorized as either close-ended (e.g., Yes/No questions) or open-ended.

Despite its promising potential, medical VQA is still in its nascent stage and current systems fall short of delivering satisfactory results. The majority of existing solutions [75, 26, 34, 29, 28] belong to the so-called closed-type category, as shown in Figure 1(a), where each potential answer is treated as a separate class. These methods apply a classification model on the fused image-question representation to predict the correct answer. The benefit of this formulation lies in its simplicity and reduced

answer space, making the task more tractable. While such methods perform well on binary or close-ended questions, they often struggle with the more complex open-ended questions, which are inherently longer and more variable.

Alternatively, a smaller body of work [1, 49] models medical VQA as a generative task, referred to as open-type approaches (illustrated in Figure 1(b)). These methods generate answers sequentially, word by word, conditioning each word on the previous ones and the input features. This design allows for deeper interaction between the image-question context and the generated answer, which can be advantageous for open-ended question scenarios. However, due to the enormous space of potential outputs, generative models frequently produce irrelevant or hallucinated answers, thereby lowering the overall accuracy. As a result, these methods are not yet widely adopted in medical VQA. Though some research efforts [89] have attempted to combine both paradigms, they typically adopt a hard separation: using classification for closed-ended questions and generation for open-ended ones.

To overcome this dichotomy and advance the field, we propose a novel architecture termed **Q2ATransformer**, which we define as a semi-open type framework, illustrated in Figure 1(c). Our semi-open approach retains the classification-based structure to maintain a constrained answer space, while simultaneously incorporating answer semantics to facilitate interactions between fused image-question representations and candidate answers—much like generative models. This hybrid strategy enables our method to mitigate the limitations of closed-type frameworks while leveraging the strengths of open-type systems.

Concretely, we introduce a set of learnable candidate answer embeddings and allow them to interact with the fused image-question features via a Transformer decoder. In this setup, the answer embeddings act as queries to assess their relevance to the input pair. Through attention mechanisms, each candidate embedding is evaluated in relation to the fused representation, allowing the model to determine the presence of

corresponding answer classes. This design enables our classification-based model to simulate semantic interaction akin to that in generation-based systems, while still benefiting from reduced prediction complexity. Compared with existing generation-based approaches, our model significantly enhances accuracy and stability.

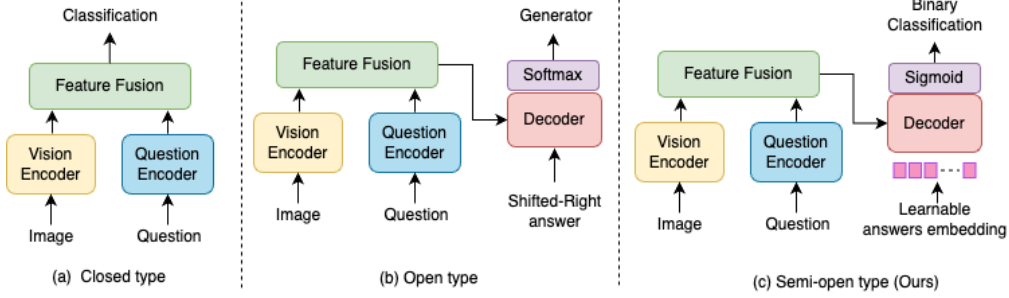
Most importantly, our model treats close-ended and open-ended questions within a unified architecture, eliminating the need for separate modules or task-specific pipelines.

The key contributions of our work are as follows:

- We introduce a novel semi-open type framework for medical VQA that blends the strengths of closed-type classification methods and open-type generative models. This is achieved via a Transformer-based architecture that leverages candidate answer embeddings while maintaining a constrained prediction space.
- We design a Cross-Modality Attention Network (CMAN) to effectively integrate image and question features. Rather than relying on element-wise multiplication or addition, CMAN concatenates the modality-specific features and employs self-attention to capture their interdependencies, leading to more informative fused representations. As demonstrated in our ablation study, CMAN outperforms conventional fusion methods in medical VQA tasks.
- Our proposed model sets new benchmarks on two widely used medical VQA datasets, delivering strong results for both close-ended and open-ended questions. In particular, it achieves substantial performance gains for open-ended questions, with absolute improvements of 16% on VQA-RAD and 41% on PathVQA, confirming the effectiveness of our semi-open framework.

## 3.2 Method

In this section, we introduce **Q2ATransformer**, a semi-open structured model tailored for medical VQA tasks. We begin with a high-level overview



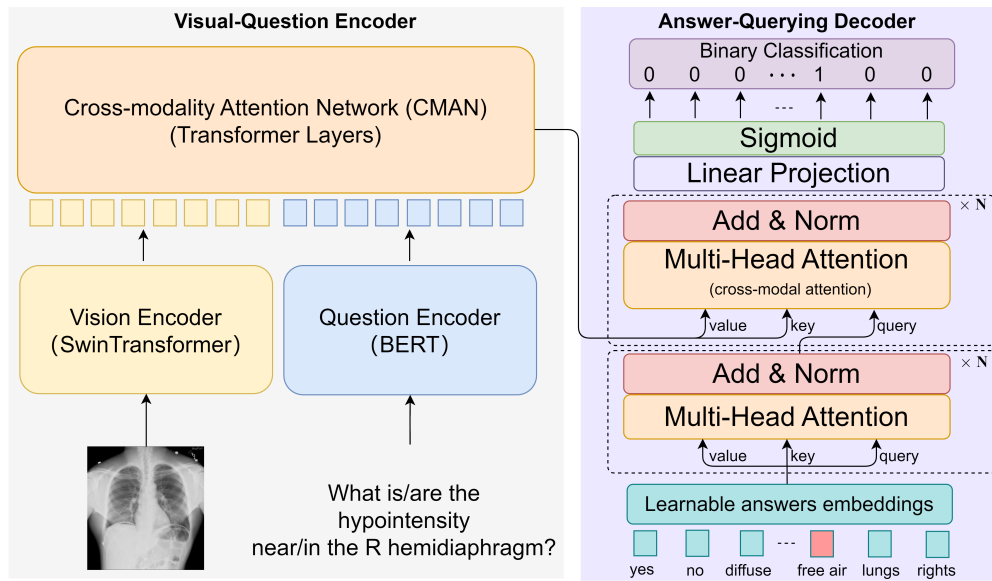
**Figure 3.1:** Different paradigm designs for medical VQA models. (a) The closed-type paradigm formulates VQA as a classification problem, where a prediction head directly operates on the joint representation of the image and question to identify the correct answer class. (b) The open-type paradigm follows a generative modeling approach, where the fused image-question features interact sequentially with previously generated answer tokens to produce the next token via a text decoder. (c) Our semi-open framework introduces a decoder to learn candidate answer embeddings, which then engage with the fused image-question features to enhance the accuracy of answer class prediction.

of the model, followed by detailed discussions of the Visual-Question Encoder in Sec. 3.1 and the Answer-Querying Decoder in Sec. 3.2.

As illustrated in Fig. 3.2, our proposed Q2ATransformer shares the general philosophy of classifying answer candidates, similar to most existing medical VQA approaches. However, it innovatively incorporates learnable candidate answer embeddings into the decision process. The architecture is composed of two main modules: a *Visual-Question Encoder* and an *Answer-Querying Decoder*.

The Visual-Question Encoder is responsible for processing the input, which includes a medical image and a clinically meaningful question. It outputs a joint representation that encodes both visual and textual modalities. This encoder is composed of three sub-components: a vision encoder, a question encoder, and a fusion module. Specifically, we adopt the Swin Transformer to encode image features and utilize BERT for question encoding. To merge the information from both modalities, we design a **Cross-Modality Attention Network (CMAN)**, which effectively combines the outputs of the two encoders into a unified feature representation.

The Answer-Querying Decoder takes this fused image-question feature and a set of learnable candidate answer embeddings as inputs, and computes the likelihood scores for each answer candidate. This decoder comprises two stacked Transformer decoder layers, followed by a classification head that predicts the final answer class probabilities.



**Figure 3.2:** Overview of the Q2ATransformer architecture. The input pair, consisting of a medical image and a corresponding question, is processed by the Visual-Question Encoder, which extracts and integrates features from both modalities. This encoder includes a Swin Transformer for visual feature extraction, a BERT model for question encoding, and a Cross-Modality Attention Network designed to effectively fuse the two feature types. The resulting fused representation is then passed into the Answer-Querying Decoder. In this stage, learnable candidate answer embeddings serve as queries in an attention mechanism, interacting with the fused features to refine their representations and assess the likelihood of each potential answer being correct.

### 3.2.1 Visual-Question Encoder

The Visual-Question Encoder is composed of three primary components: an image encoder, a question encoder, and a multimodal fusion module. Each part is detailed below.

**Image Encoder.** We adopt the Swin Transformer [70] as our image feature extractor, moving beyond traditional CNN-based architectures.

This choice is motivated by three major advantages. First, Swin Transformer structures the vision transformer hierarchically, similar to CNNs, making it adaptable to multi-scale features and maintaining linear computational complexity with increasing image size. Second, it introduces cross-window communication via window shifting, enabling the modeling of long-range dependencies and enhancing inter-patch interactions. This capability results in richer regional representations, which are particularly beneficial for the detailed textures often present in medical images. Third, the model benefits from pretraining on large-scale datasets, making it a highly robust and generalizable feature extractor.

Given an input image  $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ , where  $C$  denotes the number of channels, and  $H$  and  $W$  represent the height and width, the image embeddings  $\mathbf{F}_i \in \mathbb{R}^{N \times D_f}$  are obtained through the following transformation:

$$\mathbf{F}_i = \mathbf{W}_i \times \text{SwinTransformer}(\mathbf{I}) + \mathbf{b}_i,$$

where  $\mathbf{W}_i$  and  $\mathbf{b}_i$  are learnable parameters used to project the Swin Transformer output into a unified feature dimension  $D_f$ , aligning it with the dimension of the question embeddings. This projection also offers adaptability for fine-tuning the encoder to better suit the specific characteristics of our medical datasets. Here,  $N$  denotes the number of extracted image region features.

#### Question Encoder.

To encode the input question, we employ the pre-trained BERT model [25], which has demonstrated strong performance across a wide range of NLP tasks, including question answering. BERT integrates contextual information from both the left and right of each word, allowing it to generate rich, bidirectional representations of input sequences. Its effectiveness in capturing semantic dependencies makes it a suitable choice for our task as the question encoder.

The question embeddings  $\mathbf{F}_q \in \mathbb{R}^{M \times D_f}$  are computed as:

$$\mathbf{F}_q = \text{BERT}(\mathbf{Q}_e),$$

where  $\mathbf{Q}_e$  is the input question,  $M$  denotes the number of extracted question tokens, and  $D_f$  represents the embedding dimension.

### Feature Fusion Mechanism.

Once the image and question features are extracted, we introduce a novel fusion module—**Cross-Modality Attention Network (CMAN)**—to integrate information from the visual and textual modalities. Given the fine-grained nature of medical images and the subtle distinctions that often carry clinical significance, our design emphasizes preserving detail and enabling rich cross-modal interactions.

In CMAN, we first concatenate the image features  $\mathbf{F}_i$  and question features  $\mathbf{F}_q$  into a unified sequence:

$$\mathbf{F}_c = [\mathbf{F}_i; \mathbf{F}_q].$$

Unlike conventional fusion strategies such as matrix multiplication or element-wise summation, concatenation retains more information from both modalities and sets the stage for enhanced interaction modeling.

Next, the combined sequence  $\mathbf{F}_c$  is fed into two Transformer encoder layers, which use self-attention mechanisms to model pairwise relationships between image and question tokens. This interaction enriches the representation with cross-modal dependencies. The final fused representation  $\mathbf{F}_f$  is computed as follows:

$$\begin{aligned} \mathbf{F}_c &= [\mathbf{F}_i; \mathbf{F}_q] \\ \mathbf{Q}_{F_c} &= \mathbf{W}_q \mathbf{F}_c, \quad \mathbf{K}_{F_c} = \mathbf{W}_k \mathbf{F}_c, \quad \mathbf{V}_{F_c} = \mathbf{W}_v \mathbf{F}_c \\ \mathbf{F}_{att} &= \text{Att}(\mathbf{Q}_{F_c}, \mathbf{K}_{F_c}, \mathbf{V}_{F_c}) = \text{softmax} \left( \frac{\mathbf{Q}_{F_c} \mathbf{K}_{F_c}^T}{\sqrt{d_k}} \right) \mathbf{V}_{F_c} \\ \mathbf{F}_f &= \mathbf{W}_f \mathbf{F}_{att} + \mathbf{b}_f \end{aligned} \tag{3.1}$$

Here,  $\mathbf{W}_q$ ,  $\mathbf{W}_k$ ,  $\mathbf{W}_v$ ,  $\mathbf{W}_f$ , and  $\mathbf{b}_f$  are trainable parameters. The notation “;” denotes the concatenation operation. The matrices  $\mathbf{Q}_{F_c}$ ,  $\mathbf{K}_{F_c}$ , and  $\mathbf{V}_{F_c}$  serve as the *query*, *key*, and *value*, respectively, within the self-attention

mechanism, and are computed via linear projections of the concatenated input  $\mathbf{F}_c$ .

### 3.2.2 Answer Querying Decoder

Given an input image question pair, among a set of answers of interest, our Answer Querying Decoder predicts whether each candidate answer matches the corresponding image question pair and uses the candidate with the highest probability as the final answer. For this purpose, we employ a two-layer transformer decoder followed by a linear projector as our classifier, and introduce a set of learnable candidate answer embeddings together with the fused image-question feature  $\mathbf{F}_f$  as the input of the decoder. Assuming there are  $C$  answer classes in total, we need  $C$  candidate answer embeddings with one-to-one correspondence to the  $C$  answer classes. These answer embeddings, collectively represented by a matrix  $\mathbf{A}$ , are randomly initialised and will be updated during training through a self-attention module, a cross-attention module, and a feed-forward network(FFN) in order. Both the self-attention module and the cross-attention module implement the multi-head self-attention ( $MSA(query, key, value)$ ) but with different *key*, *query*, and *value*. The self-attention module computes the relation between different answer embeddings by using  $\mathbf{A}$  to construct all the *key*, *query*, and *value* matrices. The cross-attention module cares about the relation between the answer embeddings  $\mathbf{A}$  and the fused image-question feature  $\mathbf{F}_f$ . It thus uses the answer embedding  $\mathbf{A}$  as the *query* and the fused image-question feature  $\mathbf{F}_f$  as the *key* and *value* to compute the attention and further updates the answer embeddings by combining the attended image-question features. Mathematically, denoting the answer embeddings at the  $l$ -th layer as  $\mathbf{A}_l$ , it will be updated from the output of the previous layer  $\mathbf{A}_{l-1}$  as follows:

$$\begin{aligned} \mathbf{A}_l &= MSA(\mathbf{A}_{l-1}, \mathbf{A}_{l-1}, \mathbf{A}_{l-1}) \\ \mathbf{A}_l &= MSA(\mathbf{A}_l, \mathbf{F}_f, \mathbf{F}_f) \\ \mathbf{A}_l &= FFN(\mathbf{A}_l), \end{aligned} \tag{3.2}$$

where  $l = 1 \cdots L$  and  $L$  is the number of Transformer decoder layers. Through this process, the image-question features are injected into the answer embeddings and used to refine the latter. The refined  $C$  answer embeddings are sent to the final linear projection layer followed by a sigmoid function  $\sigma(\cdot)$  to predict the probabilities of answer classes. That is:

$$\mathbf{p} = \sigma(\mathbf{W}_A \mathbf{A}_L + \mathbf{b}), \quad (3.3)$$

where  $\mathbf{W}_A$  and  $\mathbf{b}$  are learnable parameters, and  $\mathbf{p}$  is a vector comprising of  $C$  probabilities corresponding to  $C$  answer classes. The answer class with the highest probability is chosen as the predicted answer.

### 3.2.3 Loss Function

One major challenge in medical VQA lies in the pronounced imbalance between different answer types—binary (Yes/No) responses dominate, while longer open-ended answers are significantly underrepresented. To better handle this imbalance during training, we adopt a simplified variant of the asymmetric loss, which is a modified version of focal loss. This formulation adjusts the focusing parameter  $\gamma$  asymmetrically for positive and negative samples, as described in Eqn. 3.4:

$$\mathcal{L} = \frac{1}{C} \sum_{c=1}^C \begin{cases} (1 - p_c)^{\gamma^+} \log(p_c), & \text{if } y_c = 1 \\ (p_c)^{\gamma^-} \log(1 - p_c), & \text{if } y_c = 0 \end{cases} \quad (3.4)$$

Here,  $y_c$  denotes the binary ground-truth label that specifies whether the answer class  $c$  is correct for a given image-question pair, and  $p_c$  represents the predicted probability for class  $c$ . The total loss  $\mathcal{L}$  is averaged over all answer classes and training samples. By default, we set the hyperparameters to  $\gamma^+ = 1$  and  $\gamma^- = 4$  to place more emphasis on addressing the dominance of easy negative samples.

## 3.3 Experiments And Results

### 3.3.1 Datasets

We evaluate our proposed method on two widely used medical visual question answering benchmarks: VQA-RAD [51] and PathVQA [34], which are introduced below.

**VQA-RAD** is one of the most established datasets in the domain of radiology-based VQA. It comprises 315 medical images and a total of 3,515 question-answer (QA) pairs, where each image is associated with at least one QA instance. The questions span 11 semantic categories, including “anomalies”, “properties”, “color”, “number”, “morphology”, “organ type”, “other”, and “section”. Approximately 58% of the questions are close-ended, with the remaining being open-ended. The images cover anatomical regions such as the head, chest, and abdomen. Since no official split is provided, we follow the dataset partitioning protocol outlined in the MMQ method [26] to ensure consistency with prior work.

**PathVQA** is a large-scale benchmark designed for pathology-oriented VQA tasks. It is constructed by collecting pathology images along with their captions from digital educational resources, such as electronic textbooks and online archives. The dataset contains 32,799 QA pairs in total. Among these, open-ended questions constitute 50.2% of the dataset. The closed-ended questions are well balanced, with 8,145 labeled “yes” and 8,189 labeled “no”. PathVQA includes 1,670 images sourced from two pathology textbooks and 3,328 additional images retrieved from the PEIR digital library [83]. For fair comparison, we adopt the same training and testing splits as in the MMQ setting [26].

### 3.3.2 Comparison with the State-of-the-Art Methods

We compare our proposed Q2ATransformer with seven leading medical VQA models: StAn [34], BiAn [34], MAML [29], MEVF [75], MMQ [26], PubMedCLIP [28], and MMBERT [49]. Among these, the first six methods are classification-based and are selected for their strong performance on the VQA-RAD and PathVQA benchmarks. MMBERT [49] is included

as a representative generation-based approach, with reported results on VQA-RAD.

For fairness, we adopt the same data split strategy used in MMQ [26], which is also employed by PubMedCLIP [28] and MMBERT [49], making all results directly comparable. Performance scores for PubMedCLIP and MMBERT are taken from their respective original papers, while the rest are sourced from the MMQ [26] evaluation.

As presented in Table 3.1, our Q2ATransformer achieves superior performance across both datasets. On **VQA-RAD**, our model obtains an accuracy of 79.19% on open-ended questions, representing an absolute improvement of 16.09% over the second-best method. For close-ended questions, we reach 81.2%, exceeding previous results by 1.2%. Across all question types, Q2ATransformer yields an overall accuracy of 80.48%, outperforming the strongest baseline by 8.48%.

On **PathVQA**, our model continues to deliver robust results. We achieve 54.85% accuracy on open-ended questions (a 41.45% absolute improvement), 88.85% on Yes/No questions (a 4.85% gain), and 74.61% overall (25.81% improvement). Further performance gains are possible by increasing the dimensionality of the candidate answer embeddings, as demonstrated in our ablation studies.

These results clearly highlight the advantage of Q2ATransformer, particularly in handling open-ended questions. This validates our design intuition—enabling interaction between fused image-question features and answer semantics facilitates better understanding and prediction for complex answers. Moreover, our method surpasses the generation-based MMBERT [49], benefiting from a reduced answer space that mitigates the generation of irrelevant or non-existent answers.

### 3.3.3 Ablation Study

To better understand the contributions of our proposed Cross-Modality Attention Network (CMAN) and the Answer-Querying Decoder, we conduct a series of ablation experiments under different model configurations, as summarized in Table 3.2. Specifically, we compare three fusion strategies for integrating image and question features: Bilinear Attention

**Table 3.1:** Performance comparison of different methods. † and ‡ indicate the methods are classification-based(closed-type) or generation-based(open-type), respectively.

| References<br>Methods        | Fusion<br>Methods | PathVQA      |              |              | VQA-RAD      |             |              |
|------------------------------|-------------------|--------------|--------------|--------------|--------------|-------------|--------------|
|                              |                   | Free-form    | Yes/No       | Over-all     | Open-ended   | Close-ended | Over-all     |
| StAn <sup>†</sup> [34]       | SAN               | 1.6          | 59.4         | 30.5         | 24.2         | 57.2        | 44.2         |
| BiAn <sup>†</sup> [34]       | BAN               | 2.9          | 68.2         | 35.6         | 28.4         | 67.9        | 52.3         |
| MAML <sup>†</sup> [29]       | SAN               | 5.4          | 75.3         | 40.5         | 38.2         | 69.7        | 57.1         |
|                              | BAN               | 5.9          | 79.5         | 42.9         | 40.1         | 72.4        | 59.6         |
| MEVF <sup>†</sup> [75]       | SAN               | 6.0          | 81.0         | 43.6         | 40.7         | 74.1        | 60.7         |
|                              | BAN               | 8.1          | 81.4         | 44.8         | 43.9         | 75.1        | 62.7         |
| MMQ <sup>†</sup> [26]        | SAN               | 11.2         | 82.7         | 47.1         | 46.3         | 75.7        | 64.0         |
|                              | BAN               | 13.4         | 84.0         | 48.8         | 53.7         | 75.8        | 67.0         |
| PubMedCLIP <sup>†</sup> [28] | -                 | -            | -            | -            | 60.1         | 80          | 72.1         |
| MMBERT <sup>†</sup> [49]     | -                 | -            | -            | -            | 63.1         | 77.9        | 72.0         |
| Ours                         |                   | <b>54.85</b> | <b>88.85</b> | <b>74.61</b> | <b>79.19</b> | <b>81.2</b> | <b>80.48</b> |

**Table 3.2:** Ablation Studies. BAN, SAN, and CMAN stand for Bilinear Attention Network [50], Stacked Attention Network [118] and ours Cross-modality Attention Network, respectively; Decoder refers to our Answer-Querying Decoder.

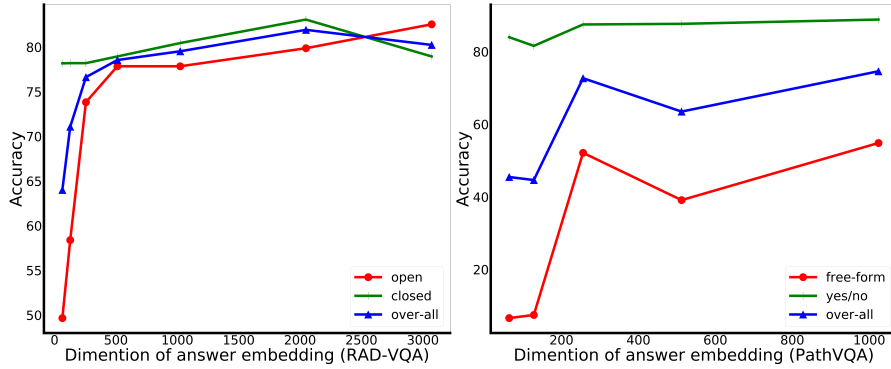
| # | BAN | SAN | CMAN | Decoder | VQA-RAD |        |         | PathVQA   |        |         |
|---|-----|-----|------|---------|---------|--------|---------|-----------|--------|---------|
|   |     |     |      |         | open    | closed | overall | free-form | yes/no | overall |
| 1 | ✓   |     |      |         | 43.62   | 75.56  | 64.1    | 15.03     | 78.24  | 51.69   |
| 2 | ✓   |     |      | ✓       | 54.36   | 80.07  | 70.84   | 44.78     | 88.29  | 70.09   |
| 3 |     | ✓   |      |         | 61.07   | 77.07  | 71.33   | 44.58     | 86.29  | 68.88   |
| 4 |     | ✓   |      | ✓       | 73.83   | 80.08  | 77.83   | 52.88     | 88.44  | 73.51   |
| 5 |     |     | ✓    |         | 69.13   | 76.32  | 73.73   | 47.53     | 86.73  | 70.31   |
| 6 |     |     | ✓    | ✓       | 79.19   | 81.2   | 80.48   | 54.85     | 88.85  | 74.61   |

Network (BAN) [50], Stacked Attention Network (SAN) [118], and our CMAN. The presence or absence of the Answer-Querying Decoder is also evaluated. All experiments are conducted with identical vision and language encoders to ensure fairness. The ✓ symbol denotes that a specific component is included in the configuration.

**Effectiveness of CMAN.** The performance gains observed when moving from configuration #1 to #5 or from #3 to #5 clearly demonstrate the superiority of our CMAN module over BAN and SAN. Unlike BAN, which fuses modalities via element-wise multiplication, or SAN, which applies direct matrix summation, CMAN concatenates the features from both modalities and then applies Transformer-based attention for interaction. This approach significantly reduces the risk of information loss introduced by more aggressive operations such as multiplication or summation, thereby yielding more expressive fused features.

**Contribution of the Answer-Querying Decoder.** Our decoder consistently enhances model performance across different attention fusion strategies. As shown in Table 3.2, adding the decoder (compare #2 vs. #1, #4 vs. #3, and #6 vs. #5) leads to noticeable performance improvements regardless of the base fusion mechanism. This indicates that the decoder is robust and complementary. Notably, when combined with CMAN (configuration #6), our model achieves the best performance, validating the synergy between CMAN and the decoder design.

**Influence of Answer Embedding Dimension.** We further investigate how the dimensionality of the candidate answer embeddings affects model performance, with results shown in Fig. 3. The model’s accuracy improves as the embedding size increases, reaching its peak around 2048 dimensions. However, larger embeddings also introduce higher computational costs, and performance gains begin to plateau. Considering the trade-off between efficiency and accuracy, we adopt a 1024-dimensional setting for answer embeddings in our final model configuration.



**Figure 3.3:** Ablation study about different dimensions of answer embeddings.

### 3.3.4 Qualitative Results

We present representative prediction examples from the PathVQA and VQA-RAD datasets in Figure 3.4 and Figure 3.5, respectively. As illustrated, our Q2ATransformer model—when equipped solely with the proposed CMAN fusion module (i.e., without the decoder)—is able to correct the majority of the errors made by MMQ, which employs the BAN mechanism for feature fusion. These results highlight the effectiveness of our CMAN in capturing richer cross-modal interactions.

Moreover, incorporating the Answer-Querying Decoder leads to further improvements. By leveraging learnable candidate answer embeddings that interact with the fused image-question features, our full model demonstrates enhanced capability in accurately selecting the correct answers, particularly in complex cases where simple fusion alone may fall short.

As shown in Figure 3.6, the model’s attention is primarily concentrated on the image and question features that are closely associated with the predicted answer, highlighting its ability to align visual and textual cues with the final response.

### 3.3.5 Limitation and Discussion

As outlined in Section 3.2, our model represents each answer class as a learnable embedding, and all such embeddings are used as queries in

the Answer-Querying Decoder to compute attention scores. While this design enables rich interactions between candidate answers and fused image-question features, it also introduces a scalability issue: the global self-attention mechanism incurs increased computational cost when the number of answer classes becomes very large.

This limitation is analogous to challenges encountered in NLP tasks involving long input sequences, where full attention becomes computationally expensive. To mitigate this, several strategies have been proposed, such as sparse or dynamic attention mechanisms, which approximate the attention map by selectively focusing on a subset of tokens. Inspired by such advances, future work will explore the integration of sparse attention techniques to improve the scalability and efficiency of our decoder without compromising its predictive performance.

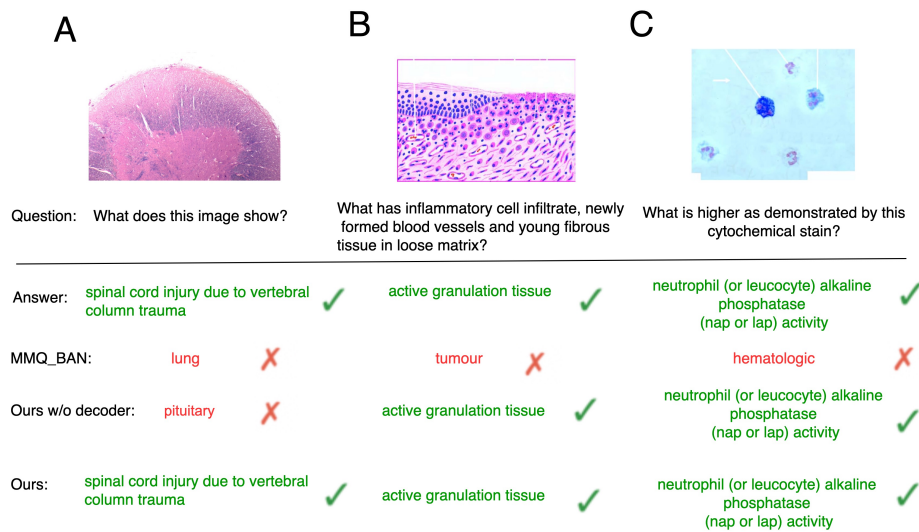


Figure 3.4: Example results from PathVQA dataset.




|                   | A   | B   | C   |
|-------------------|---|---|---|
| Image             |  |  |  |
| Question:         | Where are the infarcts?   | What skeletal joint is seen in this image?  | Where are the signals most intense?   |
| Answer:           | basal ganglia, cerebellum<br>cerebral cortex ✓                                    | Sacroiliac joint ✓  | frontal and occipital ✓   |
| MMQ_BAN:          | diffuse ✗   | right subdural hematoma ✗   | lateral ventricles  |
| Ours w/o decoder: | diffuse ✗   | sacroiliac joint ✓  | frontal and occipital ✓   |
| Ours:             | basal ganglia, cerebellum<br>cerebral cortex ✓                                    | sacroiliac joint ✓  | frontal and occipital ✓   |

Figure 3.5: Example results from VQA-RAD dataset.

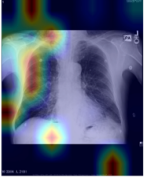

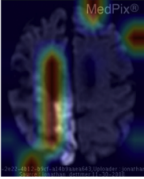
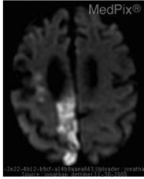
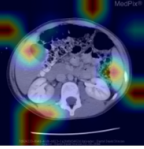

|   |   |   |
|---|---|---|
|  |  | <u>Question:</u> “Is free air present under the patient's left or right hemidiaphragm?” |
|   |   | <u>Answer:</u> “Right”  |
|  |  | <u>Question:</u> “Where is the brain tissue ischemic?”                                  |
|   |   | <u>Answer:</u> “Right upper lobe”   |
|  |  | <u>Question:</u> “What is/are the hypodense mass near/in the liver?”                    |
|   |   | <u>Answer:</u> “Gallbladder”  |

Figure 3.6: Examples of model attention focusing on informative regions in the image and question. The visualizations highlight how the model aligns visual and textual features to generate the corresponding answer.



## Chapter 4

# SAT-RRG: Self-Adaptive Training for Radiology Report Generation Leveraging LLMs for Dynamic Token-Level Refinement

In this chapter, we examine the limitations of current radiology report generation (RRG) systems and explore strategies to enhance model performance. Existing RRG approaches are primarily optimized using cross-entropy (CE) loss, which rewards only exact token matches with the ground-truth reports. This rigid supervision neglects semantically valid variations and fails to identify or correct factual inconsistencies in generated reports. Furthermore, conventional training pipelines lack mechanisms for semantic-level feedback, treating all tokens equally and overlooking the importance of correcting critical errors. To address these shortcomings, we propose **SAT-RRG**, a self-adaptive training framework designed to improve factual consistency and semantic precision in radiology report generation. SAT-RRG leverages the capabilities of the underlying large language model (LLM) within the RRG system to autonomously detect semantically incorrect tokens during training—without the need for human-annotated error labels. Guided by these pseudo-labeled error spans, we introduce two token-level objectives: **CTAL** (Confident Token Adaptive Loss), which reinforces tokens predicted with high confidence and correctness, and **ETAPL** (Error Token Adaptive Penalty Loss), which

discourages overconfident but incorrect semantic predictions. This dual-objective supervision enables the model to adaptively focus its learning: consolidating reliable content while correcting factual errors. Built upon a unified LLM backbone that supports both report generation and error identification, SAT-RRG introduces no additional computational overhead during inference and requires no manual annotations. Empirical results demonstrate that SAT-RRG achieves new state-of-the-art performance on two benchmark datasets—MIMIC-CXR and IU-Xray.

## 4.1 Introduction

Radiology report generation (RRG) has emerged as a promising direction to alleviate the workload of radiologists and enhance diagnostic consistency. Recent progress in deep learning, especially through encoder-decoder frameworks and large language models (LLMs) [98, 17, 16, 103, 39, 10], has demonstrated the feasibility of automatically generating descriptive reports from medical images. These models typically generate text in an autoregressive manner—predicting one token at a time—by modeling each token as a classification problem with corresponding probability distributions. Training is generally conducted using cross-entropy (CE) loss, which maximizes the likelihood of selecting ground-truth tokens based on the reference reports.

However, CE loss enforces strict token-level alignment, which makes the model overly sensitive to exact word choices, often penalizing semantically equivalent alternatives. Furthermore, most existing RRG methods lack mechanisms to evaluate and integrate semantic consistency during training—specifically, whether the generated text is coherent and clinically accurate in relation to the reference—let alone utilize such semantic feedback to adjust learning behavior.

In natural language generation (NLG), prior work has explored token-level feedback using reinforcement learning [129], contrastive learning [62], or post-hoc correction strategies. Yet, these approaches often depend on handcrafted reward functions or require additional labeled

data, making them impractical for the RRG setting due to the high cost of expert annotations.

To overcome these limitations, we propose a self-adaptive training framework for RRG that introduces dynamic token-level supervision without relying on external supervision signals. Rather than incorporating auxiliary reward models or annotations, our method leverages the inherent semantic reasoning capabilities of the LLM already embedded in the RRG architecture. This allows the model to autonomously identify semantic inconsistencies between its own predictions and the ground-truth report during training.

Concretely, we design an internal LLM-based mechanism to detect semantically incorrect tokens by comparing generated reports against reference reports. Unlike prior methods that depend on static reference alignment, our approach enables token-level probability adjustments that prioritize the correction of factual errors while maintaining fluency in correct predictions. To support this adaptive training scheme, we introduce two novel loss functions: **Correct Token Augment Loss (CTAL)**, which reduces the gradient updates for confidently correct tokens to encourage stability; and **Error Token Adaptive Penalty Loss (ETAPL)**, which intensifies the penalty on high-confidence but incorrect tokens, promoting targeted error correction. Together, these objectives guide the model toward learning more robust and clinically consistent representations.

Our key contributions are as follows:

**First**, we propose a self-adaptive RRG framework that leverages the built-in semantic understanding of LLMs to dynamically guide token-level optimization. This approach enables precise error correction while preserving the fluency and coherence of report generation. Unlike previous methods focused on strict word alignment, our strategy emphasizes phrase-level semantic correctness.

**Second**, we introduce two specialized loss functions, CTAL and ETAPL, that serve complementary roles in optimizing the model. CTAL stabilizes learning by reinforcing correct predictions, whereas ETAPL selectively penalizes semantically incorrect tokens predicted with undue confidence.

**Third**, our framework achieves new state-of-the-art results on two widely used RRG benchmarks: MIMIC-CXR and IU-Xray. Importantly, it does so without adding inference-time computational burden, as the same LLM backbone is used for both generation and semantic error detection.

## 4.2 Related Work

**Radiology Report Generation.** The task of radiology report generation (RRG) focuses on automatically producing clinically meaningful textual interpretations from medical images. Traditional encoder-decoder architectures [17, 66] typically rely on word-level cross-entropy (CE) loss, which constrains semantic flexibility and penalizes medically valid paraphrasing. Recent studies have sought to address these limitations by incorporating domain-specific knowledge or architectural enhancements. For example, METransformer [104] employs expert-derived tokens to refine cross-modal attention mechanisms; EKAGen [10] integrates instance-level clinical knowledge to guide attention supervision; and KiUT [39] introduces U-shaped modality connections to improve clinical alignment. However, these models largely lack fine-grained, token-level feedback during training and treat all output tokens equally—without distinguishing between semantically correct and incorrect content. *Our approach addresses this shortcoming by moving beyond rigid token-level supervision, enabling the model to recognize and retain semantically correct tokens even when surface forms differ, thereby promoting more meaning-preserving learning.*

**LLM-Based Report Generation.** Large language models (LLMs) such as LLaMA [96], Qwen [114], and their medical adaptations [105, 64] have shown significant promise for report generation. These methods typically align visual representations with LLM token embeddings to enable free-form generation. For instance, R2GenGPT utilizes a linear mapper to project image features into the LLM’s input space, while Bootstrapping leverages in-domain prompt engineering and a coarse-to-fine decoding strategy to improve factuality. Despite these advances, most LLM-based RRG models do not incorporate token-level semantic feedback or error-aware supervision during training. *In contrast, our method introduces a*

*lightweight error-tagging mechanism that enables precise token-level feedback without the need for manual labels, allowing the model to selectively refine incorrect predictions.*

**Token-Level Supervision and Feedback Learning.** Fine-grained supervision at the token level has been explored in general natural language generation through techniques such as reinforcement learning [129], contrastive learning objectives [62], and correction-based fine-tuning. However, these approaches often rely on handcrafted reward signals or require additional human-annotated data, which limits their scalability, particularly in specialized domains such as medical NLP. Consequently, token-level supervision remains underutilized in RRG due to the cost and complexity of expert annotation. *To mitigate this, we employ in-context prompting with a frozen LLM to automatically annotate semantically inconsistent tokens in generated reports using special tags (e.g., <e>...</e>). This self-supervised error identification process allows for scalable training without reliance on human labeling. Furthermore, we introduce a dual-objective loss framework: the **Error Token Adaptive Penalty Loss (ETAPL)**, which penalizes overconfident but incorrect tokens, and the **Correct Token Augment Loss (CTAL)**, which reinforces confidently correct predictions. CTAL acts as a stabilizing force, while ETAPL provides targeted correction, together enabling a more adaptive and clinically aware learning process.*

### 4.3 Methodology

An overview of the proposed SAT-RRG framework is shown in Figure 4.1. The framework is composed of three major components: (a) *report generation*, described in Section 4.3.1; (b) *error token identification*, detailed in Section 4.3.2; and (c) the *inference process*, illustrated in the right portion of Figure 4.1.

The overall training procedure is formally presented in Algorithm 1, and we include a concrete step-by-step example in Section 4.3.4 to facilitate understanding.

---

**Algorithm 1** Training Phase of SAT-RRG

---

**Input:** Chest X-ray image  $X_{\text{img}}$ , Prompt  $P$ , Ground truth report  $R_{\text{GT}}$

**Output:** Trainable loss  $L_{\text{total}}$

**1 Step 1: Input Construction**

Extract visual features:  $Z_v \leftarrow f_{\text{img}}(X_{\text{img}})$   
 Project to embedding space:  $H_v \leftarrow \text{MLP}(Z_v)$   
 Tokenize:  $P_{\text{tok}} \leftarrow \text{Tokenizer}(P)$ ,  $R_{\text{GT\_tok}} \leftarrow \text{Tokenizer}(R_{\text{GT}})$   
 Concatenate:  $X \leftarrow \text{Concat}(H_v, P_{\text{tok}}, R_{\text{GT\_tok}})$   
 Build causal attention mask  $M$  over  $R_{\text{GT\_tok}}$

**2 Step 2: Forward Pass**

$\hat{y} \leftarrow \text{LLM.forward}(X, \text{mask} = M)$

**3 Step 3: Self-Critique Error Detection**

Generate decoded prediction:  $y \leftarrow \text{Argmax}(\hat{y})$   
 Detect semantic errors:  
 $\text{error\_tokens} \leftarrow \text{LLM.detect}(R_{\text{GT}}, y)$   
 $T_{\text{correct}} \leftarrow \{y_j \mid y_j \notin \text{error\_tokens}\}$   
 $T_{\text{error}} \leftarrow \{y_k \mid y_k \in \text{error\_tokens}\}$

**4 Step 4: Loss Computation**

$\mathcal{L}_{\text{correct}} \leftarrow -\sum_{j \in T_{\text{correct}}} p(y_j) \log p(y_j)$   
 $\mathcal{L}_{\text{error}} \leftarrow -\sum_{k \in T_{\text{error}}} p(y_k) \log p(y_k)$   
 $\mathcal{L}_{\text{GT}} \leftarrow -\sum_{i=1}^T y_i^{\text{GT}} \log p(\hat{y}_i)$   
 Total loss:  $\mathcal{L}_{\text{total}} \leftarrow \lambda(\mathcal{L}_{\text{correct}} + \mathcal{L}_{\text{error}}) + \mathcal{L}_{\text{GT}}$

**5 Step 5: Backpropagation**

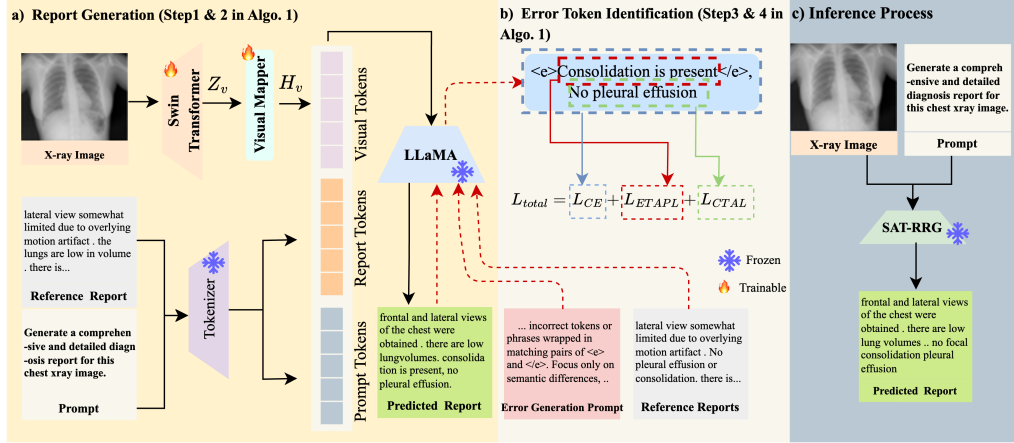
Update model using  $\nabla_{\theta} \mathcal{L}_{\text{total}}$

---

### 4.3.1 Report Generation

The report generation module in SAT-RRG comprises three main components: (1) a visual encoder that extracts features from the input radiograph, (2) a visual mapper that projects these features into the word embedding space of the LLM, and (3) a large language model (LLM) that generates diagnostic reports conditioned on both visual and textual inputs.

Given a chest X-ray image  $X_v$ , we utilize a Swin Transformer [70] to extract high-level visual representations, yielding feature maps  $Z_v = \text{Swin}(X_v)$ . These features encode salient regions relevant for clinical interpretation. To enable compatibility with the LLM’s embedding space, we introduce a multi-layer perceptron (MLP) as a visual mapper, which



**Figure 4.1:** Overview of the SAT-RRG framework. It includes three components: (a) Report generation process, (b) Error token identification via LLM-based semantic comparison, and (c) Inference process. Tokens or modules with the same color represent the same entity across different stages. After the generator produces the initial predicted report, it is passed through the LLM again for feedback. At this stage, each token in the predicted report retains its original generation probability, since the model has not yet been updated by any loss. The LLM-based comparison only annotates which tokens are semantically incorrect. These annotations serve as token-level supervision targets: incorrect tokens are penalized and correct tokens reinforced via two complementary losses (ETAPL and CTAL). The original probabilities are used by the losses to propagate training signals back into the model.

transforms the image features into LLM-aligned embeddings:  $H_v = \text{MLP}(Z_v)$ .

The transformed visual features  $H_v$  are then concatenated with a textual prompt  $P$  and the ground-truth report  $R_{\text{GT}}$ , both of which are tokenized for LLM processing. The complete input to the LLM is constructed as:

$$X = \text{Concat}(H_v, \text{Tokenizer}(P), \text{Tokenizer}(R_{\text{GT}})),$$

where  $P$  serves as a guiding instruction, such as: “Generate a comprehensive and detailed diagnosis report for this chest X-ray image.”

The LLM generates a sequence of report tokens  $y = \{y_1, y_2, \dots, y_T\}$ , which are supervised using the standard word-level cross-entropy loss during training.

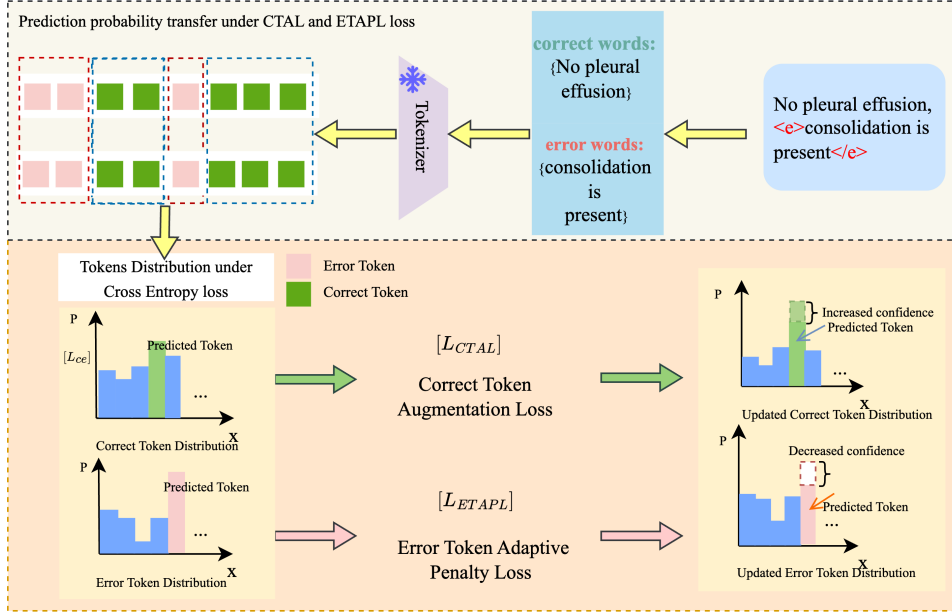
### 4.3.2 Dynamic Error Token Identification and Correction

Beyond the core report generation process, SAT-RRG incorporates a dynamic error correction mechanism that leverages the LLM’s semantic reasoning capability to identify and rectify phrase-level inconsistencies between generated and reference reports. As illustrated in Figure 4.1(b), this component classifies each token in the generated report as either semantically correct or incorrect based on its consistency with the ground truth.

To achieve this, we prompt the LLaMA3 model with both the generated and reference reports and instruct it to annotate semantically incorrect tokens using custom error tags (`<e> . . . </e>`), while leaving correct tokens unmarked. The prompt includes explicit instructions and rules to guide the LLM in identifying clinically significant semantic discrepancies and distinguishing them from permissible paraphrases. We adopt a few-shot prompting strategy—providing labeled examples within the prompt—to help the model generalize the error annotation behavior effectively. Detailed prompt design is discussed in Section A, and representative annotation examples are shown in Figure 4.4.

As demonstrated in Figure 4.4, only tokens with clear semantic contradictions are annotated. For instance, if the generated report contains "right lower lobe pneumonia" while the ground truth states "no evidence of pneumonia," the phrase is marked as erroneous. Similarly, a generated statement of "no pleural effusion" contradicts the ground truth description of "small right and moderate left pleural effusion," leading to corresponding error annotations. In contrast, semantically equivalent expressions—such as "no evidence of pneumonia" and "there is no focal consolidation concerning for pneumonia"—are not marked, reflecting their clinical alignment despite surface-level differences. This token-level supervision enables the model to focus on true semantic errors rather than superficial lexical mismatches.

By distinguishing correct from incorrect tokens, SAT-RRG enables the application of targeted loss functions to improve model training, as



**Figure 4.2:** Overview of our proposed training framework with token-level supervision. During training, the predicted report is dynamically compared to the reference report using the LLM’s semantic understanding to identify correct and erroneous tokens. Tokens are annotated with **green** for correct and **red** for incorrect cases via inline tags. The **Correct Token Augmentation Loss (CTAL)** encourages higher confidence in correct tokens by promoting gradient flattening, while the **Error Token Adaptive Penalty Loss (ETAPL)** penalizes overconfident incorrect predictions. This token-level feedback facilitates interpretable and adaptive optimization for higher-quality generation.

depicted in Figure 4.2. Semantically correct tokens are reinforced using the **Correct Token Augment Loss** ( $L_{CTAL}$ ), which increases their predicted probability and stabilizes accurate outputs. In contrast, erroneous tokens are penalized via the **Error Token Adaptive Penalty Loss** ( $L_{ETAPL}$ ), which reduces their likelihood in subsequent generations, discouraging the repetition of factual errors. The formulations of these loss functions are presented in the following section.

### 4.3.3 Error-Aware Training Objectives

Conventional cross-entropy (CE) loss promotes grammatical fluency and syntactic correctness by aligning predicted tokens with reference tokens.

However, it lacks the capacity to distinguish between minor lexical variations and critical semantic errors, treating all mismatches uniformly. To introduce more fine-grained and clinically relevant supervision, we design two auxiliary token-level losses: one that reinforces confidently correct tokens and another that penalizes high-confidence semantic errors, thereby facilitating more precise and adaptive learning.

**CTAL: Reinforcing Confidently Correct Tokens.** Let  $T_{\text{correct}}$  denote the set of tokens identified as semantically correct (i.e., those not enclosed within  $\langle e \rangle \dots \langle /e \rangle$ ). To reinforce such tokens, we define the Correct Token Augmentation Loss (CTAL) as:

$$\mathcal{L}_{\text{CTAL}} = - \sum_{j \in T_{\text{correct}}} p(y_j) \log p(y_j), \quad (4.1)$$

with its corresponding gradient:

$$\frac{\partial \mathcal{L}_{\text{CTAL}}}{\partial p(y_j)} = -\log p(y_j) - 1. \quad (4.2)$$

This formulation amplifies the probability of correct tokens, particularly those with low confidence, while attenuating updates for already confident ones. As a result, CTAL stabilizes accurate predictions and promotes semantic reliability.

**ETAPL: Penalizing Overconfident Semantic Errors.** Let  $T_{\text{error}}$  be the set of semantically incorrect tokens, enclosed by  $\langle e \rangle \dots \langle /e \rangle$ . The Error Token Adaptive Penalty Loss (ETAPL) is defined as:

$$\mathcal{L}_{\text{ETAPL}} = \sum_{k \in T_{\text{error}}} p(y_k) \log p(y_k), \quad (4.3)$$

with the gradient:

$$\frac{\partial \mathcal{L}_{\text{ETAPL}}}{\partial p(y_k)} = 1 + \log p(y_k). \quad (4.4)$$

ETAPL dynamically penalizes overconfident erroneous predictions. Tokens predicted with high probability (i.e., high  $p(y_k)$ ) receive stronger

gradient suppression, discouraging repeated semantic errors. In contrast, low-confidence mistakes are penalized more gently to avoid destabilizing the training dynamics.

**Total Loss.** The final training objective combines these two auxiliary losses with the standard cross-entropy loss  $\mathcal{L}_{\text{GT}}$ , yielding:

$$\mathcal{L}_{\text{total}} = \lambda \cdot (\mathcal{L}_{\text{CTAL}} + \mathcal{L}_{\text{ETAPL}}) + \mathcal{L}_{\text{GT}}, \quad (4.5)$$

where  $\lambda$  is a tunable coefficient that balances the influence of self-supervised semantic feedback against standard token-level supervision.

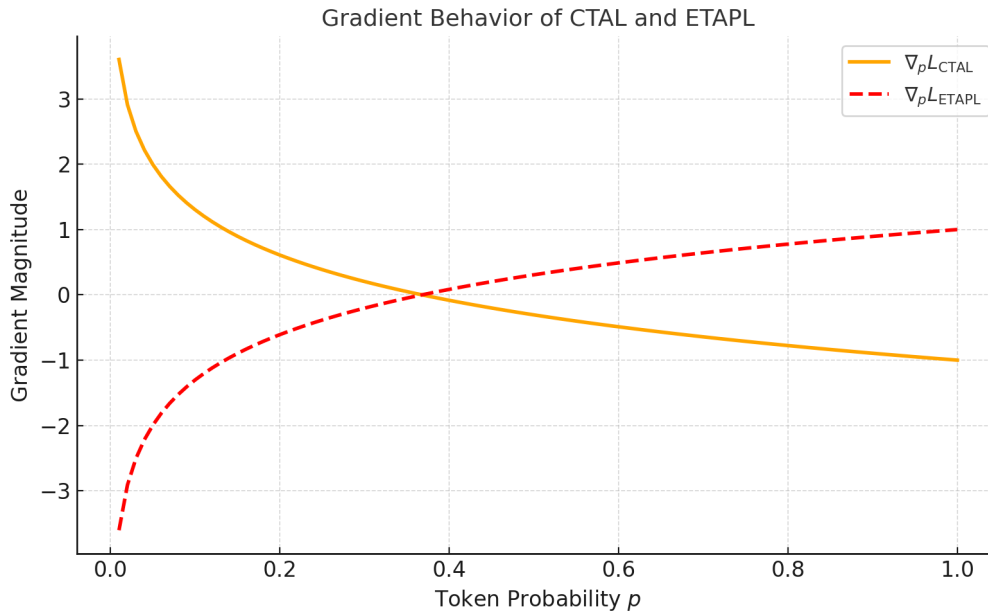
The gradient of the total objective with respect to model parameters  $\theta$  is computed as:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{total}}}{\partial \theta} = & \lambda \sum_{j \in T_{\text{correct}}} \frac{\partial p(y_j)}{\partial \theta} \cdot (-\log p(y_j) - 1) \\ & + \lambda \sum_{k \in T_{\text{error}}} \frac{\partial p(y_k)}{\partial \theta} \cdot (1 + \log p(y_k)) \\ & - \sum_{i=1}^T \frac{\partial p(y_i^{\text{pred}})}{\partial \theta} \cdot \frac{y_i^{\text{GT}}}{p(y_i^{\text{pred}})}. \end{aligned} \quad (4.6)$$

This formulation provides an interpretable and directionally guided optimization signal: it strengthens low-confidence correct predictions, suppresses high-confidence semantic errors, and maintains alignment with the reference report. Together, CTAL and ETAPL enable semantically informed token-level learning, enhancing both clinical fidelity and linguistic precision in report generation.

### Loss Formulation and Gradient Analysis

The gradients of the two proposed auxiliary loss functions—CTAL and ETAPL—offer directionally meaningful and interpretable supervision signals, as illustrated in Figure 4.3. CTAL encourages the model to increase the confidence of semantically correct predictions, particularly those with low initial probability, while ETAPL penalizes incorrect predictions made with high confidence. Together, these gradients provide



**Figure 4.3: Gradient dynamics of CTAL and ETAPL.** CTAL applies stronger gradients to low-confidence correct tokens, reinforcing their confidence during training. ETAPL imposes larger penalties on high-confidence incorrect tokens, thereby suppressing overconfident semantic errors. This complementary gradient behavior enables stable and semantically guided optimization.

adaptive feedback that promotes both training stability and semantic precision.

#### 4.3.4 Example Workflow Illustration

To illustrate the working mechanism of our self-adaptive token-level refinement framework, we present a concrete example that demonstrates the full end-to-end process. The framework operates through four key stages: (1) **Report Generation**, where the initial diagnostic report is produced by the model; (2) **Semantic Error Annotation**, where an LLM identifies semantically incorrect tokens using in-context comparison with the reference report; (3) **Loss-Based Refinement**, in which CTAL and ETAPL are selectively applied based on token correctness; and (4) **Token Probability Update**, where the model adjusts its parameters to reinforce correct predictions and suppress semantic errors.

| Dataset                     | Methods                          | @B-1         | @B-2         | @B-3         | @B-4         | METEOR       | ROUGE        |
|-----------------------------|----------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| MIMIC-CXR                   | AdaAtt [72]                      | 0.314        | 0.198        | 0.132        | 0.094        | 0.128        | 0.131        |
|                             | M2Transformer [21]               | 0.332        | 0.210        | 0.142        | 0.101        | 0.134        | 0.142        |
|                             | R2Gen <sup>†</sup> [17]          | 0.353        | 0.218        | 0.145        | 0.103        | 0.142        | -            |
|                             | R2GenCMN <sup>†</sup> [16]       | 0.353        | 0.218        | 0.148        | 0.106        | 0.142        | -            |
|                             | PPKED <sup>†</sup> [66]          | 0.360        | 0.224        | 0.149        | 0.106        | 0.149        | 0.237        |
|                             | METransformer <sup>†</sup> [104] | 0.386        | 0.250        | 0.169        | 0.124        | 0.152        | 0.291        |
|                             | DCL <sup>†</sup> [54]            | -            | -            | -            | 0.109        | 0.150        | 0.284        |
|                             | KiUT <sup>†</sup> [39]           | 0.393        | 0.243        | 0.159        | 0.113        | 0.160        | 0.285        |
|                             | R2GenGPT <sup>†</sup> [105]      | 0.411        | <u>0.267</u> | <u>0.186</u> | <u>0.134</u> | 0.160        | <u>0.297</u> |
|                             | EKAGen <sup>†</sup> [10]         | <u>0.419</u> | 0.258        | 0.170        | 0.119        | 0.157        | 0.287        |
|                             | Bootstrapping <sup>†</sup> [64]  | 0.402        | 0.262        | 0.180        | 0.128        | <b>0.175</b> | 0.291        |
|                             | CvT2DistilGPT2 <sup>†</sup> [76] | 0.393        | 0.248        | 0.171        | 0.127        | -            | 0.155        |
|                             | RaDialog-RG <sup>†</sup> [85]    | 0.346        | -            | -            | 0.095        | 0.140        | 0.271        |
|                             | Multi-Grained <sup>†</sup> [63]  | 0.346        | 0.226        | 0.159        | 0.117        | 0.163        | 0.290        |
| PromptMRG <sup>†</sup> [44] | 0.398                            | -            | -            | 0.112        | 0.157        | 0.268        |              |
| <b>Ours</b>                 | <b>0.431</b>                     | <b>0.284</b> | <b>0.199</b> | <b>0.146</b> | <u>0.168</u> | <b>0.306</b> |              |
| IU-Xray                     | AdaAtt [72]                      | 0.284        | 0.207        | 0.150        | 0.126        | 0.165        | 0.311        |
|                             | M2Transformer [21]               | 0.402        | 0.284        | 0.168        | 0.143        | 0.170        | 0.328        |
|                             | R2Gen <sup>†</sup> [17]          | 0.470        | 0.304        | 0.219        | 0.165        | 0.187        | 0.371        |
|                             | R2GenCMN <sup>†</sup> [16]       | 0.475        | 0.309        | 0.222        | 0.170        | 0.191        | 0.375        |
|                             | METransformer <sup>†</sup> [104] | 0.483        | 0.322        | 0.228        | 0.172        | 0.192        | 0.380        |
|                             | DCL <sup>†</sup> [54]            | -            | -            | -            | 0.163        | 0.193        | -            |
|                             | R2GenGPT <sup>†</sup> [105]      | 0.488        | 0.316        | 0.228        | 0.173        | <u>0.211</u> | 0.377        |
|                             | Bootstrapping <sup>†</sup> [64]  | <u>0.499</u> | 0.323        | 0.238        | 0.184        | 0.208        | 0.390        |
|                             | CvT2DistilGPT2 <sup>†</sup> [76] | 0.473        | 0.304        | 0.224        | 0.175        | 0.200        | 0.376        |
|                             | Multi-Grained <sup>†</sup> [63]  | 0.472        | 0.321        | 0.234        | 0.175        | 0.192        | 0.379        |
|                             | PromptMRG <sup>†</sup> [44]      | 0.401        | -            | -            | 0.098        | 0.160        | 0.281        |
|                             | EKAGen [10]                      | 0.497        | <u>0.339</u> | <u>0.250</u> | 0.190        | 0.210        | <u>0.399</u> |
| <b>Ours</b>                 | <b>0.504</b>                     | <b>0.342</b> | <b>0.254</b> | <b>0.196</b> | <b>0.222</b> | <b>0.400</b> |              |

**Table 4.1:** Comparison on MIMIC-CXR and IU-Xray datasets(@B stands for BLUE).

This step-by-step pipeline enables dynamic, interpretable, and fine-grained supervision during training.

**Initial Report Generation (Steps 1 & 2 in Algorithm 1).** Given an input chest X-ray image and a textual prompt, the frozen LLM generates an initial draft report:

*Prediction:* Consolidation is present, no pleural effusion.

Each token in the generated report is associated with a model-predicted confidence score (probability). For example:

| Token         | Probability |
|---------------|-------------|
| Consolidation | 0.85        |
| is            | 0.90        |
| present       | 0.92        |
| no            | 0.98        |
| pleural       | 0.96        |
| effusion      | 0.93        |

**Self-Critiqued Error Annotation (Step 3 in Algorithm 1).** The same LLM is then prompted to compare the generated report against the corresponding reference:

*Reference:* No pleural effusion or consolidation.

Using few-shot prompting and its built-in semantic reasoning ability, the LLM identifies a contradiction: the phrase “consolidation is present” conflicts with the reference. This span is annotated using inline error tags:

*Annotated:* <e>Consolidation is present</e>, no pleural effusion.

**Token-Level Supervision via CTAL and ETAPL (Step 4 in Algorithm 1).** Tokens identified as correct (no, pleural, effusion) are reinforced using the Correct Token Augmentation Loss (CTAL), which increases their predicted probabilities. In contrast, the semantically erroneous tokens (consolidation, is, present) are penalized using the Error Token Adaptive Penalty Loss (ETAPL), especially when their initial confidence is high.

**Parameter Update (Step 5 in Algorithm 1).** After applying the respective losses, model parameters are updated via backpropagation, resulting in refined token probabilities:

| Token         | Initial | Action    | Updated |
|---------------|---------|-----------|---------|
| Consolidation | 0.85    | Penalize  | 0.60    |
| is            | 0.90    | Penalize  | 0.65    |
| present       | 0.92    | Penalize  | 0.58    |
| no            | 0.98    | Reinforce | 0.99    |
| pleural       | 0.96    | Reinforce | 0.98    |
| effusion      | 0.93    | Reinforce | 0.97    |

This token-level self-correction mechanism allows the model to not only learn from the ground-truth reference but also adaptively refine itself based on its own semantic inconsistencies, enabling more accurate and fluent report generation over time.

## 4.4 Experiments

### 4.4.1 Datasets.

We evaluate our proposed SAT-RRG framework on two widely used radiology report generation benchmarks: MIMIC-CXR and IU-Xray.

#### MIMIC-CXR

[46] is a large-scale chest radiograph dataset containing 377,110 images and 227,835 associated reports collected from 64,588 patients at Beth Israel Deaconess Medical Center between 2011 and 2016. Following the partitioning protocol of Chen et al. [17], we use 270,790 images for training and 3,858 for testing to ensure fair comparisons.

#### IU-Xray

[23] is another widely adopted benchmark for RRG, comprising 3,955 radiology reports and 7,470 chest X-ray images. We adopt the same data split strategy as in [17], dividing the dataset into training, validation, and test subsets in a 7:1:2 ratio, and reporting performance on the test set.

### 4.4.2 Implementation Details.

We adopt the LLaMA3-3B model<sup>1</sup> as our large language model and employ the Swin Transformer<sup>2</sup> as the visual encoder. The loss balancing coefficient is set to  $\lambda = 0.5$ . Training is conducted on two NVIDIA A6000 GPUs (48GB each), with a mini-batch size of 24 and a learning rate of 5e-5. For inference, we apply beam search with a beam width of 3 to balance efficiency and output diversity.

### 4.4.3 Evaluation Metrics.

We evaluate the quality of generated reports using standard natural language generation metrics, including BLEU [82], METEOR [6], and ROUGE-L [61], following the official evaluation protocol<sup>3</sup>. For assessing clinical fidelity, we exclude CheXBert [40] due to prior inconsistencies in its implementation and unclear aggregation variants (e.g., micro vs. macro), which undermine fair comparison. Instead, we adopt three robust and widely validated clinical metrics: RadGraph F1 [41], BERTScore [123], and RadCliQ [120]. Notably, RadCliQ incorporates CheXBert in a standardized fashion, making it more reliable for measuring clinical alignment.

| Ground Truth Report   | Report Before TLS   | Report After TLS   |
|---|---|--|
| impression : <b>no evidence of pneumonia</b> or decompensated congestive heart failure . stable findings associated with the patients known lung malignancy . findings : pa and lateral radiograph of the chest once again demonstrate a right upper lobe mass with a fiducial marker in place as well as a right perihilar mass . this is consistent with the patients known malignancy . once again seen are a small <b>right and moderate left pleural effusion</b> with considerable left lower lobe atelectasis or consolidation | impression : <b>&lt;=&gt;right lower lobe pneumonia&lt;/=&gt;</b> . findings : there is a new opacity in the <b>&lt;=&gt;right lower lobe concerning for pneumonia&lt;/=&gt;</b> . there is <b>&lt;=&gt;no pleural effusion&lt;/=&gt;</b> or pneumothorax . cardiomeastinal silhouette and hilar contours are unremarkable . <b>&lt;=&gt;left pectoral pacemaker with leads in the right atrium and right ventricle&lt;/=&gt;</b> is unchanged . <b>&lt;=&gt;median sternotomy wires&lt;/=&gt;</b> are intact . | impression : <b>&lt;=&gt;small left pleural effusion&lt;/=&gt;</b> . findings : pa and lateral chest views were obtained with patient in upright position . analysis is performed in direct comparison with the next preceding similar study of . heart size is normal . cardiomeastinal contours are normal . there is <b>&lt;=&gt;a small left pleural effusion&lt;/=&gt;</b> . there is no pneumothorax . there is no focal consolidation concerning for pneumonia . there is no pulmonary edema . bony structures are unremarkable . |

**Figure 4.4:** Comparison of error tokens in the generated report before and after TLS. The highlighted errors are marked in corresponding colors to show the corrections made.

<sup>1</sup><https://huggingface.co/meta-llama/Llama-3.2-3B>

<sup>2</sup><https://huggingface.co/microsoft/swin-base-patch4-window7-224>

<sup>3</sup><https://github.com/tylin/coco-caption>

| Methods                       | RadGraph F1 ( $\uparrow$ ) | Bert Score ( $\uparrow$ ) | RadCliQ ( $\downarrow$ ) |
|-------------------------------|----------------------------|---------------------------|--------------------------|
| R2Gen [17]                    | 0.172                      | 0.406                     | 1.228                    |
| R2GenCMN [16]                 | 0.182                      | <u>0.418</u>              | 1.182                    |
| CvT2DistilGPT2 [76]           | <u>0.196</u>               | 0.374                     | 1.220                    |
| RaDialog-RG <sup>†</sup> [85] | -                          | 0.400                     | -                        |
| R2GenGPT [105]                | 0.187                      | 0.415                     | 1.207                    |
| PromptMRG [44]                | 0.190                      | 0.357                     | <u>1.169</u>             |
| EKAGen <sup>††</sup> [10]     | 0.199                      | 0.412                     | 1.126                    |
| <b>Ours</b>                   | <b>0.205</b>               | <b>0.422</b>              | <b>1.150</b>             |

Note: EKAGen uses 300×300 images while the others use 224×224 images. We exclude CheXBert due to inconsistencies in prior evaluations and lack of transparency, which hinder fair comparison. Notably, RadCliQ integrates CheXBert in a standardized manner.

**Table 4.2:** Evaluation of Clinic-related Metrics on MIMIC-CXR.

| Dataset   | $L_{\text{error}}^{\text{penalty}}$ | $L_{\text{correct}}$ | @B-1  | @B-2  | @B-3  | @B-4  | METEOR | ROUGE_L |
|-----------|-------------------------------------|----------------------|-------|-------|-------|-------|--------|---------|
| Mimic-cxr |                                     |                      | 0.423 | 0.273 | 0.188 | 0.135 | 0.166  | 0.295   |
|           | ✓                                   |                      | 0.426 | 0.277 | 0.192 | 0.139 | 0.166  | 0.298   |
|           |                                     | ✓                    | 0.428 | 0.281 | 0.196 | 0.141 | 0.167  | 0.303   |
|           | ✓                                   | ✓                    | 0.431 | 0.284 | 0.199 | 0.146 | 0.168  | 0.306   |

**Table 4.3:** Ablation study for loss components.

#### 4.4.4 Main Results.

We compare our proposed method with state-of-the-art (SOTA) baselines on both the MIMIC-CXR and IU-Xray datasets. The results of standard NLG metrics are reported in Table 4.1, while clinical relevance scores are shown in Table 4.2. Methods marked with † are quoted directly from their original publications, whereas all other results are reproduced using publicly available code under the same training-testing partition as our method.

As shown in Table 4.1, our approach consistently outperforms prior methods across nearly all evaluation metrics on both datasets. This includes traditional image captioning models [72, 21], transformer-based encoder-decoder architectures [17, 16, 66, 106, 39, 10], and LLM-based generation methods [105, 64, 85].

In particular, our method surpasses the performance of large LLM-based models such as R2GenGPT [105] and Bootstrapping [64] in BLEU

scores by 7.5% and 12.5%, respectively—even though those models employ larger 7B-scale LLMs, while ours uses a smaller 3B variant.

Additionally, knowledge-enhanced models like EKAGen [10] and KiUT [39], although incorporating domain-specific information, lack token-level feedback mechanisms, limiting their ability to make fine-grained corrections. Multi-Grained [63], which uses sentence-level contrastive learning, similarly falls short due to the absence of token-level supervision. Notably, EKAGen employs 300×300 resolution images instead of the standard 224×224, making direct comparison less fair; nonetheless, our method still outperforms it.

On the IU-Xray dataset, our model maintains strong and consistent performance. In addition to achieving high scores on NLG metrics, it also leads on clinical relevance metrics—RadGraph F1, BERTScore, and RadCliQ—demonstrating the benefit of our self-adaptive training strategy. The ability to dynamically provide token-level supervision enables the model to reinforce correct predictions and penalize semantic errors, contributing to both textual fluency and clinical accuracy.

#### 4.4.5 Ablation Study.

##### loss components

Table 4.3 presents the results of our ablation study on the proposed loss components. We evaluate the contribution of each loss term by selectively enabling the **Correct Token Augmentation Loss (CTAL)**, denoted as  $L_{\text{correct}}$ , and the **Error Token Adaptive Penalty Loss (ETAPL)**, denoted as  $L_{\text{error}}$ . Both components independently improve performance compared to the baseline model, which excludes token-level supervision. Notably, the best results are achieved when both CTAL and ETAPL are applied jointly, demonstrating their complementary nature in refining semantic correctness and report quality.

Table 4.5 presents clinical metric comparisons. Methods marked with † are quoted from their original papers, while others were re-run using publicly released code on the same training-test partition as ours.

### Clinical analysis results across all 14 defined categories

Table 4.4 provides a disease-wise comparison between our model and the baseline across multiple lexical, semantic, and clinical metrics. Overall, our method achieves **consistent improvements across nearly all disease categories**, demonstrating strong robustness in capturing both textual quality and clinical accuracy. Notably, diseases with complex or subtle radiographic manifestations—such as *Atelectasis*, *Edema*, *Consolidation*, and *Pleural Effusion*—show clear gains in BLEU, ROUGE-L, and METEOR scores.

Semantic and clinical metrics (BERTScore, RadCliQ, and RadGraph) further reinforce this trend, with our model producing more clinically faithful entity relationships and structured findings. Importantly, improvements extend beyond common diseases: challenging or less frequent categories, including *Pneumothorax*, *Support Devices*, and *Cardiomegaly*, also exhibit measurable gains.

These results highlight that our fine-grained reward alignment enables **better generalization across heterogeneous clinical conditions**, improving both narrative accuracy and structured clinical fidelity.

**Table 4.4:** Disease-wise comparison between our model and the baseline across multiple metrics.

| Disease         | Bleu_1      | Bleu_2      | Bleu_3      | Bleu_4      | ROUGE_L     | METEOR      | BERT_Score  | RadCliQ     | RadGraph    |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Atelectasis     | 0.26 / 0.23 | 0.15 / 0.13 | 0.09 / 0.08 | 0.06 / 0.05 | 0.17 / 0.16 | 0.14 / 0.12 | 0.27 / 0.28 | 1.41 / 1.42 | 0.19 / 0.19 |
| Cardiomegaly    | 0.24 / 0.21 | 0.14 / 0.11 | 0.08 / 0.07 | 0.05 / 0.04 | 0.16 / 0.15 | 0.14 / 0.12 | 0.26 / 0.27 | 1.43 / 1.44 | 0.18 / 0.18 |
| Consolidation   | 0.24 / 0.19 | 0.13 / 0.10 | 0.08 / 0.06 | 0.05 / 0.04 | 0.15 / 0.14 | 0.14 / 0.11 | 0.25 / 0.24 | 1.55 / 1.55 | 0.20 / 0.17 |
| Edema           | 0.25 / 0.20 | 0.13 / 0.10 | 0.08 / 0.06 | 0.05 / 0.03 | 0.16 / 0.14 | 0.15 / 0.13 | 0.27 / 0.25 | 1.39 / 1.45 | 0.20 / 0.17 |
| Fracture        | 0.21 / 0.18 | 0.12 / 0.09 | 0.06 / 0.05 | 0.04 / 0.03 | 0.15 / 0.14 | 0.13 / 0.12 | 0.25 / 0.22 | 1.55 / 1.57 | 0.18 / 0.18 |
| Pneumonia       | 0.23 / 0.15 | 0.12 / 0.07 | 0.06 / 0.03 | 0.04 / 0.01 | 0.14 / 0.13 | 0.12 / 0.10 | 0.22 / 0.20 | 1.57 / 1.65 | 0.18 / 0.15 |
| Pneumothorax    | 0.23 / 0.15 | 0.12 / 0.07 | 0.06 / 0.03 | 0.04 / 0.01 | 0.16 / 0.12 | 0.12 / 0.10 | 0.24 / 0.21 | 1.54 / 1.65 | 0.16 / 0.15 |
| LungLesion      | 0.19 / 0.12 | 0.10 / 0.06 | 0.05 / 0.03 | 0.03 / 0.01 | 0.13 / 0.10 | 0.10 / 0.08 | 0.20 / 0.17 | 1.54 / 1.61 | 0.15 / 0.13 |
| NoFinding       | 0.27 / 0.19 | 0.17 / 0.10 | 0.12 / 0.08 | 0.10 / 0.05 | 0.22 / 0.21 | 0.24 / 0.22 | 0.42 / 0.42 | 1.71 / 1.70 | 0.34 / 0.31 |
| PleuralEffusion | 0.26 / 0.21 | 0.13 / 0.10 | 0.09 / 0.07 | 0.06 / 0.04 | 0.16 / 0.15 | 0.14 / 0.13 | 0.27 / 0.24 | 1.39 / 1.43 | 0.20 / 0.19 |
| PleuralOther    | 0.26 / 0.22 | 0.14 / 0.11 | 0.09 / 0.07 | 0.06 / 0.04 | 0.17 / 0.14 | 0.14 / 0.12 | 0.24 / 0.22 | 1.66 / 1.67 | 0.17 / 0.16 |
| SupportDevices  | 0.26 / 0.22 | 0.13 / 0.10 | 0.08 / 0.06 | 0.05 / 0.03 | 0.15 / 0.14 | 0.13 / 0.13 | 0.26 / 0.25 | 1.61 / 1.50 | 0.17 / 0.18 |
| EnlargedCardior | 0.22 / 0.22 | 0.11 / 0.13 | 0.06 / 0.08 | 0.04 / 0.05 | 0.15 / 0.16 | 0.13 / 0.14 | 0.24 / 0.26 | 1.61 / 1.50 | 0.17 / 0.18 |
| AirspaceOpacity | 0.22 / 0.17 | 0.11 / 0.09 | 0.07 / 0.05 | 0.04 / 0.03 | 0.15 / 0.13 | 0.12 / 0.11 | 0.22 / 0.21 | 1.59 / 1.63 | 0.17 / 0.16 |

**Table 4.5:** Clinical metric comparison on the MIMIC-CXR dataset. We evaluate a range of recent radiology report generation models using RadGraph F1 ( $RG_{F1}$ ), BERTScore (BERT), RadCliQ (CliQ), GREEN, and the composite RaTEScore(RaTE). Higher values indicate better performance except for CliQ. Best and second-best results are highlighted in **bold** and underline. EKAGen is shaded to indicate that it uses higher-resolution images ( $300 \times 300$ ) than the other methods.

| Methods                       | $RG_{F1}(\uparrow)$ | BERT( $\uparrow$ ) | CliQ( $\downarrow$ ) | GREEN( $\uparrow$ ) | RaTE( $\uparrow$ ) |
|-------------------------------|---------------------|--------------------|----------------------|---------------------|--------------------|
| R2Gen [17]                    | 0.172               | 0.406              | 1.228                | 0.276               | 0.526              |
| R2GenCMN [16]                 | 0.182               | 0.418              | 1.182                | 0.297               | 0.538              |
| CvT2DistilGPT2 [76]           | 0.196               | 0.374              | 1.220                | <b>0.320</b>        | 0.527              |
| RaDialog-RG <sup>+</sup> [85] | -                   | 0.400              | -                    | -                   | -                  |
| PromptMRG [44]                | 0.190               | 0.357              | <u>1.169</u>         | 0.287               | 0.528              |
| R2GenGPT [105]                | 0.187               | 0.415              | <u>1.207</u>         | 0.300               | 0.528              |
| KARGEN [57]                   | <u>0.203</u>        | <u>0.421</u>       | 1.165                | 0.308               | 0.533              |
| EKAGen <sup>++</sup> [10]     | 0.199               | 0.412              | 1.126                | 0.256               | 0.512              |
| <b>Ours</b>                   | <b>0.205</b>        | <b>0.422</b>       | <b>1.150</b>         | <u>0.310</u>        | <b>0.540</b>       |

*Note.* EKAGen uses  $300 \times 300$  images while the others use  $224 \times 224$  images. We exclude CheXBert due to inconsistencies in prior evaluations and lack of transparency, which hinder fair comparison. Notably, RadCliQ integrates CheXBert in a standardized manner.

### Qualitative Analysis

Figure 4.4 illustrates qualitative examples that highlight the effect of token-level supervision (TLS) on report refinement. The comparison is made between reports generated before and after applying TLS. Prior to incorporating TLS, the model produces semantically incorrect or clinically irrelevant content, for example, stating “right lower lobe concerning for pneumonia” despite the reference report explicitly indicating “no evidence of pneumonia.” After enabling TLS, these errors are largely corrected. In the “Report After TLS,” the erroneous mention of “right lower lobe pneumonia” is removed, aligning the generated output with the reference in both meaning and phrasing. To aid visual interpretation, colored highlights are used to indicate corresponding clinical concepts between the reference, erroneous, and corrected reports. These examples demonstrate how TLS enables the model to selectively reinforce accurate information while suppressing semantic inconsistencies.

## Chapter 5

# A Systematic Evaluation of GPT-4V's Multimodal Capability for Chest X-ray Image Analysis

In this chapter, we conduct a comprehensive investigation into GPT-4V's capabilities in the domain of medical image analysis. Specifically, we evaluate its multimodal performance across three representative tasks: radiology report generation, medical visual question answering (VQA), and medical visual grounding. For each task, carefully designed prompts are used to elicit GPT-4V's underlying capabilities and generate high-quality responses. To achieve a thorough evaluation, we adopt three complementary approaches: quantitative analysis, human evaluation, and qualitative case studies. Our findings indicate that GPT-4V demonstrates strong comprehension of medical images, showing promising results in radiology report generation and medical VQA. However, its performance in medical visual grounding remains suboptimal and requires significant improvement. Furthermore, we observe a noticeable discrepancy between the outcomes of automated quantitative evaluation and human assessment. This inconsistency highlights the limitations of existing automatic metrics in effectively capturing the nuanced capabilities of large multimodal models such as GPT-4V, thereby underscoring the need for more robust and semantically aligned evaluation frameworks in the medical domain.

## 5.1 Introduction

**Large Language Models (LLMs)** have recently demonstrated remarkable capabilities across a wide range of domains and tasks [96, 77, 2]. The pursuit of enhancing LLMs with visual understanding has catalyzed the development of **Large Multimodal Models (LMMs)** [119, 53, 3], enabling joint reasoning over visual and textual inputs.

Among these, GPT-4V extends the GPT-4 architecture by incorporating visual perception capabilities via training on large-scale multimodal corpora. It has achieved state-of-the-art performance in image comprehension and text generation. Recent studies have explored GPT-4V’s potential on general visual tasks [112, 117], revealing its impressive capabilities through case-based qualitative analyses.

In the medical domain, LLMs and LMMs hold substantial promise for transforming clinical workflows [105, 91]. With the integration of a visual module, GPT-4V opens new avenues for automated medical image understanding. While several works have begun exploring its use in medical imaging [110], these investigations are primarily limited to case studies, lacking systematic or task-specific evaluation.

In this work, we move beyond anecdotal evidence to conduct a comprehensive evaluation of GPT-4V on three representative multimodal tasks in medical image analysis: (1) **radiology report generation**, (2) **medical visual question answering (VQA)**, and (3) **medical visual grounding**. These tasks collectively reflect GPT-4V’s ability to interpret, describe, and localize medical concepts in visual data. To ensure robust assessment, we adopt a three-pronged evaluation approach: (i) **quantitative analysis** using standard automatic metrics, (ii) **human evaluation** conducted by medical experts, and (iii) **case studies** for qualitative insights. This multi-perspective evaluation allows us to holistically analyze the strengths and limitations of GPT-4V in clinical settings.

We explore various **prompting strategies** tailored to each task. For radiology report generation, we compare zero-shot and few-shot prompting using exemplar reports. For medical VQA, few-shot prompting is

applied to align answer format with benchmark ground truths. For medical visual grounding, task-specific instructions guide GPT-4V to generate bounding box annotations. These settings are designed to activate relevant capabilities and reveal GPT-4V’s sensitivity to prompt engineering in the medical context.

**Our key findings are summarized as follows:**

- **Radiology Report Generation:** GPT-4V achieves competitive performance compared to specialized models, as measured by both linguistic and clinical metrics. Notably, human evaluations indicate that GPT-4V’s generated reports exhibit high factuality, coherence, and clinical relevance—often outperforming traditional models under human scrutiny.
- **Medical Visual Question Answering:** GPT-4V demonstrates strong reasoning abilities and generates informative answers. However, due to its flexible answer style and the rigid ground-truth formats in standard benchmarks, its performance appears modest under conventional metrics. Human evaluations reveal higher accuracy and medical validity than automated scores suggest.
- **Medical Visual Grounding:** GPT-4V underperforms in this task, struggling to generate accurate bounding boxes for clinically relevant findings. This mirrors existing limitations observed in general-domain grounding tasks and points to an important direction for future improvement.
- **Metric-Human Discrepancy:** Our evaluations show notable divergence between metric-based and human judgments. This highlights the inadequacy of current metrics in fully capturing the clinical correctness and semantic quality of GPT-4V’s outputs, underscoring the need for developing LLM-aligned evaluation protocols for medical vision-language tasks.

**Paper Outline.** The remainder of this chapter is organized as follows: Section 2 is the related work for GPT-4V. Section 3 introduces our Evaluation, which contains evaluation tasks, evaluation method, evaluation

datasets, and evaluation results. Section 4 describes our summary and discussion. We conclude the paper and discuss the limitations in Section 5.

## 5.2 Related Works

**GPT-4V** [77], also known as GPT-4 with Vision, is a state-of-the-art multimodal system developed by OpenAI that extends the capabilities of large language models (LLMs) by incorporating image understanding. This integration marks a significant advancement in the field of artificial intelligence, enabling the model to perform vision-language reasoning, multimodal content generation, and complex cross-modal tasks. Built on the GPT-4 architecture, GPT-4V is trained on large-scale paired image-text data from both internet sources and licensed datasets, positioning it as a general-purpose vision-language interface.

The model introduces novel capabilities and limitations that emerge from the joint processing of text and image inputs. It demonstrates advanced multimodal reasoning abilities across a diverse set of topics, including document understanding, scene interpretation, and grounded generation, opening new possibilities for human-AI interaction.

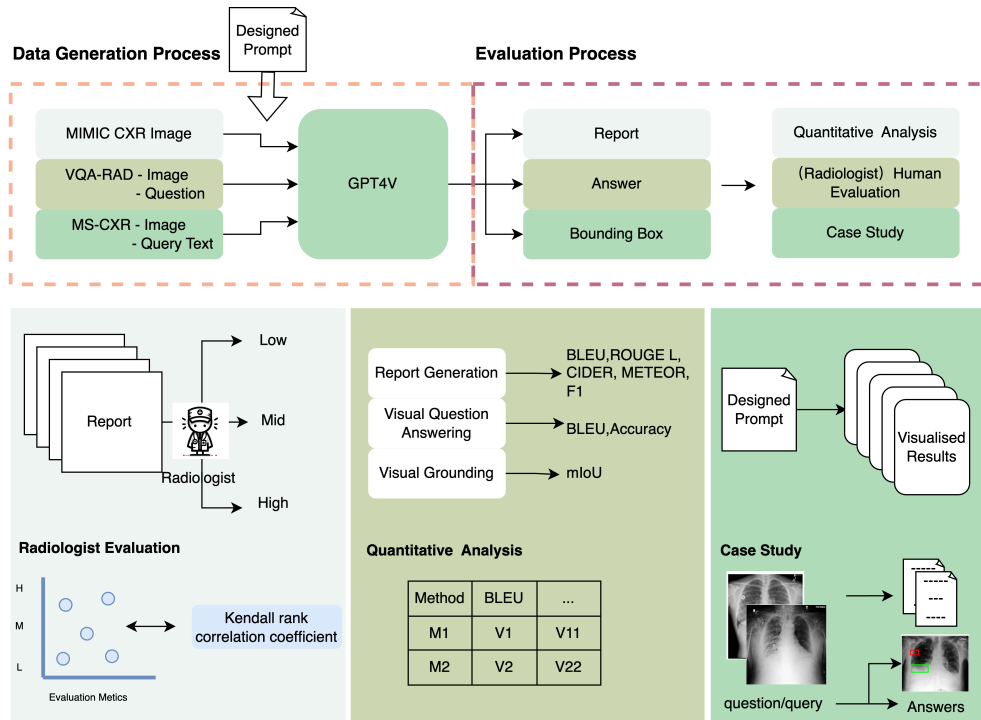
Several recent studies have examined GPT-4V's performance on general-purpose multimodal tasks. For example, Wu et al. [112] provided early insights into GPT-4V's capabilities and limitations through a qualitative case study. Li et al. [56] evaluated its performance on Visual Question Answering (VQA) benchmarks, and Shi et al. [90] investigated its Optical Character Recognition (OCR) accuracy. While these works reveal GPT-4V's potential across visual domains, they primarily focus on generic datasets and lack domain-specific evaluations.

In the medical domain, Wu et al. [110] initiated exploration of GPT-4V's application to medical images, highlighting its promise through a series of case studies. However, these analyses are largely anecdotal and do not offer systematic evaluations across tasks or datasets. In contrast, our work provides a comprehensive and task-specific evaluation

of GPT-4V’s capabilities in the medical imaging domain, including radiology report generation, medical visual question answering, and visual grounding. By incorporating both quantitative and expert-based human evaluations, we aim to uncover the strengths, limitations, and clinical viability of GPT-4V in real-world healthcare scenarios.

### 5.3 Methodology and Experiments

Figure 5.1 shows the entire evaluation process, including criteria definition, prompt design, data generation, data evaluation, and the main evaluation process.



**Figure 5.1:** The main evaluation process, depicted in the figure, encompasses the data generation process in the upper left, the evaluation process in the upper right, and, in the lower sections, the definition of criteria, doctor evaluation, analysis of NLP metrics, and case analysis.

The upcoming paragraphs provide a detailed introduction to four key components of our evaluation process.

**Evaluation Tasks** describes the specific tasks that we set up for the evaluation.

**Evaluation Method** outlines how we design the prompts and describes the methodology used to assess the models. This includes the criteria for evaluation, the metrics such as accuracy, precision, recall, F1 score, BLEU, CIDEr, ROUGE, METEOR, accuracy, and mIoU, and any specific protocols or procedures that we followed during the evaluation process.

**Evaluation Datasets** provides details about the datasets involved in the evaluation. Important aspects like the source, the sizes, and the characteristics of the datasets, as well as their relevance to the evaluation tasks, are discussed.

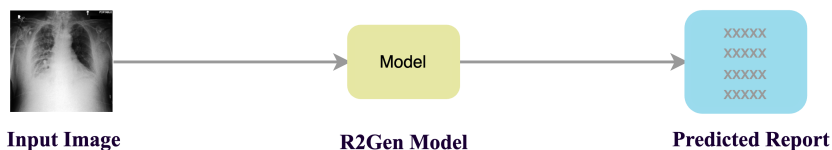
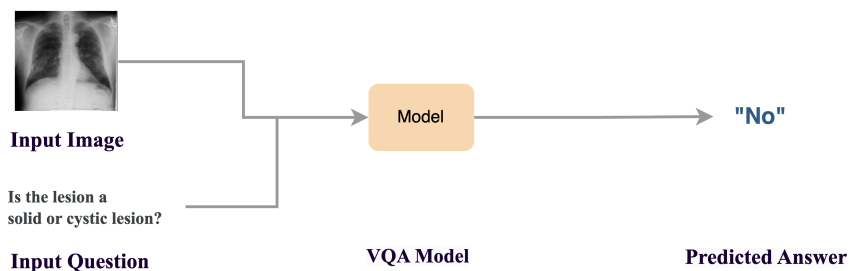
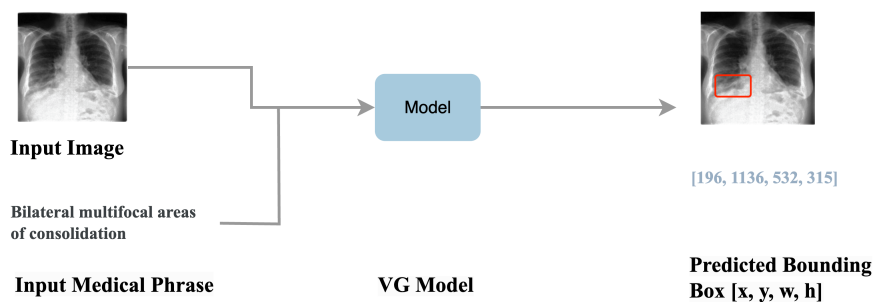
**Evaluation Results** presents the results of the evaluation. It includes a summary of the findings, comparisons with benchmarks or previous models (if applicable), and insights or interpretations of the results. It is partitioned into three parts: the quantitative, human evaluation, and case study results.

### 5.3.1 Evaluation Tasks

In the following, we introduce the three evaluation tasks shown in Figure 5.2, including radiology report generation (R2Gen), medical Visual Question Answering (VQA), and medical visual grounding (VG).

#### Radiology Report Generation

Radiology report generation (R2Gen) is a very important application of medical images, akin to the task of image captioning [101, 113, 80]. R2Gen poses unique challenges, given the inherent complexity of medical reports, their length, and the difficulty in discerning fine-grained abnormalities from medical images, particularly in datasets biased towards normal samples (referred to as the data bias problem). Current research can be grouped into two primary research directions. The first direction concentrates on enhancing the model's architecture to facilitate improved extraction of visual features and the generation of high-quality medical reports. For example, [58] used a hierarchical architecture to

**Task A: Radiology Report Generation****Task B: Medical Visual Question Answering****Task C: Medical Visual Grounding**

**Figure 5.2:** Three multimodal medical imaging tasks we employ to evaluate GPT-4V's performance.

generate reports with normality and abnormality respectively. Building on the Transformer's success [99], [17] introduced a Transformer-based model, enhancing it with relational memory and memory-driven conditional layer normalization to enhance image feature representation and capture crucial report patterns [17]. The second research direction addresses the data bias problem by incorporating external knowledge information. For example, some works constructed predefined medical knowledge graphs to augment the model's ability to capture valuable clinical information [125, 66, 54, 39].

Furthermore, very recently, there has been a surge in radiology report generation methods leveraging Large Language Models (LLMs).

These approaches harness the capabilities of LLM to generate long-text content and utilize abundant knowledge sources to enhance the quality of radiology reports. For example, [105] employed LLaMA2 [96] to elevate the quality of the generated reports, ensuring effective image-text alignment through a visual mapper.

### **Visual Question Answering**

The Visual Question Answering (VQA) task [43, 111] involves processing the input image-question pairs to generate appropriate answers. Currently, there are two predominant approaches for implementing VQA tasks: classification-based [75, 29, 28], and generation-based [1, 49]. By nature, VQA should be based on generation. The ongoing shift from a classification-centric paradigm to a generation-oriented approach represents a prevailing trend in the VQA field. Based on the characteristics of the VQA task, a proficient text generation model is essential. Consequently, the current surge in LLMs presents a significant opportunity for substantial improvements in the VQA task. Numerous endeavors incorporating LLMs into VQA tasks are already underway, whether for generating VQA datasets [84] or utilizing LLMs to enhance the performance of VQA systems [52]. The evident improvements that LLMs bring to VQA lead us to believe that they are well-suited for this task. Consequently, evaluating the VQA task is a crucial aspect of GPT-4V's evaluation.

### **Visual Grounding**

In the visual grounding (VG) [47] task, the input typically comprises a medical image accompanied by a descriptive statement about the image, often pertaining to a specific medical sign or symptom. The task's output is the coordinates of a bounding box that visually marks the area described in the statement, such as encapsulating a particular medical sign. Most visual grounding research focuses on general images, with only a few studies targeting medical images, likely due to the scarcity of corresponding medical datasets. However, the recently introduced MS-CXR dataset has opened up new possibilities in medical visual grounding,

leading to emerging publications [38, 94, 95] based on this dataset. Despite growing recognition, there remains untapped potential, presenting a significant opportunity for future research in medical visual grounding. Unlike the previous two tasks, the output for visual grounding is not a conventional text paragraph but a set of coordinates. Recent studies have successfully integrated LLMs [86, 127] to directly produce these coordinates as outputs, showing promising results. Recognizing this potential, we hypothesize that GPT-4V should also possess visual grounding capabilities. Consequently, we include this task in our evaluations to assess GPT-4V's performance in this specific area.

### 5.3.2 Evaluation Method/Process

#### Radiology Report Generation

To further harness the potential of GPT-4V, we investigate multiple prompt engineering strategies, particularly focusing on the **zero-shot** and **few-shot** paradigms.

**Zero-shot Prompt:** In the zero-shot configuration, GPT-4V receives no reference report and instead relies entirely on its pretrained knowledge to generate outputs. In this setup, GPT-4V autonomously produces both the "impression" and "findings" components, which are later compared against the corresponding ground-truth report for evaluation.

**Few-shot Prompts:** Few-shot in-context learning has emerged as a powerful technique for enhancing the performance of large language models (LLMs) [97, 107, 22]. By presenting a handful of examples within the prompt, the model can infer the desired output structure without requiring any parameter updates during inference. Unlike fine-tuning, this approach allows the model to adapt behaviorally through context alone. In our study, we explore several prompt design strategies tailored to GPT-4V under the few-shot setting. Specifically, we investigate the following configurations:

- Prompts composed solely of normal cases (**Few-shot normal-example prompt**),

- Prompts composed solely of abnormal cases (**Few-shot abnormal-example prompt**), and
- Prompts including one normal and one abnormal case (**Few-shot mixed-example prompt**).

**R2Gen prompt**

**Zero-shot prompt**  
You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. You should only generate reports and no other unnecessary information.

**Few-shot normal examples prompt**  
You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. You should only generate reports and no other unnecessary information. The format of the output report should be similar to the example:  
Examples:  
=====

*Normal Example Report 1*  
*Normal Example Report 2*

**Few-shot abnormal examples prompt**  
You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. You should only generate reports and no other unnecessary information. The format of the output report should be similar to the example:  
Examples:  
=====

*Abnormal Example Report 1*  
*Abnormal Example Report 2*

**Few-shot mixed examples prompt**  
You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. You should only generate reports and no other unnecessary information. The format of the output report should be similar to the example:  
Examples:  
=====

*Abnormal Example Report 1*  
*Normal Example Report 1*

**Figure 5.3:** R2Gen Prompt Examples. Three types of prompt settings are tested, including one zero-shot prompt and three few-shot prompts.

Our experiments indicate that incorporating both normal and abnormal example reports within prompts consistently enhances the quality of the generated outputs. Further insights into this observation can be found in our case study section. Detailed descriptions of the example reports used are provided in B.1.2. Based on this observation, we adopt the **few-shot mixed-example prompt** configuration to assess GPT-4V on the MIMIC-CXR benchmark [45]. In our evaluation, we concentrate on the **zero-shot**

and **few-shot** paradigms, intentionally excluding more intricate prompting techniques such as chain-of-thought reasoning [108] or ensembling approaches [102], in order to maintain a manageable experimental scope.

**Evaluation Metrics:** To evaluate the quality of the generated radiology reports, we employ a suite of standard NLP metrics, including BLEU, ROUGE, METEOR, and CIDEr. These metrics each offer unique insights into different facets of text generation quality. BLEU emphasizes precision by measuring the overlap between generated and reference n-grams. ROUGE, on the other hand, focuses on recall, assessing how thoroughly the generated text covers the content of the references. METEOR balances both precision and recall while incorporating semantic flexibility such as synonym handling and paraphrase detection. CIDEr, originally proposed for image captioning, evaluates how closely the generated report aligns with multiple reference texts by emphasizing consensus and weighting informative n-grams.

After computing these automatic metrics, we further conduct human evaluations to assess the alignment between model-generated reports and human expectations. Expert reviewers assess reports across multiple dimensions—relevance, factual accuracy, coherence, and overall appropriateness. Combining automatic metrics with expert human feedback enables a more comprehensive evaluation. While BLEU, ROUGE, METEOR, and CIDEr provide quantifiable measurements of surface-level similarities, human assessments capture deeper contextual, semantic, and clinical appropriateness that automatic metrics may miss. By integrating both quantitative and qualitative evaluations, we ensure a robust assessment framework that validates not only the linguistic accuracy but also the clinical validity and interpretability of generated reports.

### Medical Visual Question Answering

In the context of medical VQA, we adopt a *few-shot prompting* approach to mitigate a core limitation of current VQA datasets—their constrained predefined answer sets. To better align GPT-4V’s responses with dataset

expectations, we introduce a strategy that familiarizes GPT-4V with categorizing questions based on their anticipated answer types. Specifically, this involves distinguishing between *closed-ended* questions, which typically require short, definitive answers (e.g., “yes” or “no”), and *open-ended* questions that call for more descriptive and elaborated responses.

By incorporating this categorization into our prompting strategy, we guide GPT-4V to tailor its answers with greater precision. For closed-ended queries, the model learns to deliver succinct and direct responses. Conversely, for open-ended prompts, it generates more comprehensive explanations. This fine-grained control over response style substantially enhances both the appropriateness and informativeness of GPT-4V’s outputs, thereby improving the overall quality and reliability of our VQA evaluations.

**VQA Prompt:** Our VQA prompting strategy adheres to the template illustrated in Figure 5.4. We include seven representative examples to guide GPT-4V in producing responses that are consistent with the dataset’s expected format. These in-context examples play a crucial role in constraining the model’s output. In the absence of such guidance, GPT-4V tends to generate more verbose or unconstrained responses, which hinders the comparability between its predictions and the ground-truth answers.

**Evaluation Metrics:** To assess GPT-4V’s performance on the medical VQA task, we adopt distinct evaluation strategies for close-ended and open-ended questions.

For *close-ended questions*, we compute prediction accuracy and benchmark GPT-4V’s performance against state-of-the-art (SOTA) VQA methods, which are predominantly classification-based. This comparison highlights GPT-4V’s standing within the current landscape of medical VQA technologies, particularly in scenarios requiring concise and definitive answers (e.g., “yes” or “no”).

For *open-ended questions*, we report BLEU-4 scores, a standard metric in natural language processing that evaluates the overlap between generated and reference text. This choice reflects the inherently diverse and descriptive nature of open-ended questions, which allow for multiple plausible responses differing in phrasing and structure.

To complement the quantitative evaluation, we conduct human assessment to examine the alignment between BLEU scores and expert judgment. A board-certified radiologist reviews GPT-4V’s responses for clinical accuracy, relevance, and coherence. While BLEU offers an objective textual similarity measure, it may overlook semantic nuances or domain-specific reasoning. Thus, human evaluation plays a critical role in validating the real-world applicability of the model’s answers.

By integrating automated scoring with expert human assessment, we establish a comprehensive and robust framework for evaluating the model’s capabilities across different question types.

### **Medical Visual Grounding:**

For the Medical Visual Grounding (VG) task, we employ a straightforward prompt without incorporating examples. We opt for this approach to prevent potential restrictions on GPT-4V’s response to only those scenarios covered in the examples, thus preserving its ability to generalize. Our goal is to avoid steering the model towards producing results solely based on the provided examples. In our prompt design, we focus solely on informing GPT-4V about the required output format, i.e., the coordinates of the bounding box. This approach guides the model in understanding our specific requirements, particularly in generating bounding boxes for visual grounding tasks.

**Visual Grounding Prompt:** We construct a tailored prompt format aimed at enhancing the model’s comprehension and precision in generating bounding boxes. An example of this prompt design is presented in Figure 5.5.

**Evaluation Metrics:** We evaluate by calculating the mean Intersection over Union (mIoU) across all samples in the benchmark dataset. The mIoU metric is a standard measure used in the field to assess the accuracy of object localization and segmentation in images. By computing the mIoU, we could gauge how precisely the GPT-4V identifies and outlines relevant areas in the images compared to the ground truth. After calculating the mIoU, we conduct a comparative analysis with the SOTA VG methods. This comparison allows us to benchmark GPT-4V's performance against the currently leading methods in the field. By comparing the mIoU results of GPT-4V with those of the SOTA methods, we could gain a clear understanding of the strengths and weaknesses of GPT-4V in the context of the broader research landscape.

### 5.3.3 Evaluation Datasets

**Radiology Report Generation Dataset:** MIMIC-CXR [45] is the largest publicly available dataset in this domain, comprising chest radiographs paired with corresponding unstructured textual reports. It contains 377,110 chest X-ray images and 227,835 associated reports, collected from 64,588 patients who underwent examinations at the Beth Israel Deaconess Medical Center between 2011 and 2016. To ensure fair and consistent benchmarking, we adopt the official dataset split provided by MIMIC-CXR, resulting in a test set with 3,858 samples. For human evaluation, we randomly select 100 report pairs—each consisting of a ground-truth report and its corresponding GPT-4V-generated version—and have them annotated by a board-certified radiologist.

**VQA Dataset:** VQA-RAD [51] is one of the most widely used datasets for medical visual question answering. It comprises 315 radiology images and 3,515 question-answer (QA) pairs, with each image associated with at least one QA pair. The questions span 11 categories, including “anomalies,” “properties,” “color,” “number,” “morphology,” “organ type,” “other,” and “section.” Notably, 58% of the questions are close-ended, while the remainder are open-ended. The dataset predominantly features images from the head, chest, and abdomen regions. Following

the official split, the training set includes 3,064 QA pairs, and the test set contains 451 pairs. For human evaluation, 100 open-ended QA pairs are randomly selected from the test set and assessed by a trained radiologist.

**Visual Grounding Dataset: MS-CXR** [9] is a specialized dataset for biomedical vision–language research. It consists of 1,162 image–sentence pairs annotated with bounding boxes and corresponding textual phrases, all curated by board-certified radiologists. The dataset covers eight cardiopulmonary radiological findings with roughly balanced sample sizes. MS-CXR offers two types of bounding box annotations: reviewed edits and manually drawn boxes from scratch. A distinguishing feature of MS-CXR lies in its emphasis on complex semantic modeling and real-world clinical language understanding. It presents significant challenges for joint image–text reasoning, including the interpretation of location references, complex negations, and stylistic variability in radiology reporting. The dataset serves as a benchmark for phrase grounding and has been instrumental in demonstrating the utility of principled textual semantic modeling for enhancing self-supervised vision–language frameworks.

### 5.3.4 Evaluation Result

#### Quantitative Results

##### Radiology Report Generation

Comparison with SOTA Methods: Table 5.1 presents a performance comparison between the GPT-4V model and SOTA methods using the MIMIC-CXR dataset [45]. The comparison methods encompass standard image captioning techniques, including Show-Tell [101], Att2in [113], AdaAtt [71], Transformer [99], and M2Transformer [20]. Additionally, we compare with the radiology report generation methods, specifically R2Gen [17], R2GenCMN [16], MSAT [106], and METransformer [104]. As aforementioned, we employ few-shot mixed-example prompts to help GPT-4V generate medical reports. From Table 5.1, it is clear that radiology report generation models such as METransformer, MSAT, and R2Gen exhibit

**Table 5.1:** Comparison on the MIMIC-CXR dataset.

| Methods             | BLEU-1       | BLEU-2       | BLEU-3       | BLEU-4       | ROUGE        | METEOR       | CIDEr        |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| M2Transformer [20]  | 0.332        | 0.210        | 0.142        | 0.101        | 0.264        | 0.134        | 0.142        |
| R2Gen [17]          | 0.353        | 0.218        | 0.145        | 0.103        | 0.277        | 0.142        | -            |
| R2GenCMN [16]       | 0.353        | 0.218        | 0.148        | 0.106        | 0.278        | 0.142        | -            |
| PPKED [65]          | 0.360        | 0.224        | 0.149        | 0.106        | 0.284        | 0.149        | 0.237        |
| GSK [116]           | 0.363        | 0.228        | 0.156        | 0.115        | 0.284        | -            | 0.203        |
| MSAT [106]          | 0.373        | 0.235        | 0.162        | 0.120        | 0.282        | 0.143        | 0.299        |
| METransformer [104] | 0.386        | 0.250        | 0.169        | 0.124        | 0.291        | 0.152        | <b>0.362</b> |
| KiUT [76]           | 0.393        | 0.234        | 0.159        | 0.113        | 0.160        | 0.285        | -            |
| R2GenGPT [105]      | <b>0.411</b> | <b>0.267</b> | <b>0.186</b> | <b>0.134</b> | <b>0.297</b> | <b>0.160</b> | 0.269        |
| GPT-4V [77]         | 0.338        | 0.190        | 0.109        | 0.061        | 0.240        | 0.125        | 0.033        |

top-tier performance. Nevertheless, GPT-4V’s capability to generate medical reports is impressive, considering it is designed as a general-purpose model. Leveraging the advantages of an extensive dataset for pretraining, GPT-4V performs well in several metrics, including BLEU [81], ROUGE [60], and METEOR [5]. However, when compared to models specifically trained on MIMIC-CXR, GPT-4V exhibits a performance gap, particularly evident in CIDEr [100]. This discrepancy arises because the CIDEr metric differently scores words based on their occurrence frequencies, potentially affecting GPT-4V’s performance when it fails to generate certain MIMIC-CXR-specific words, yielding relatively lower scores. Our evaluation reveals that GPT-4V possesses the capacity to generate information not present in the ground truth but visually evident in the image. This phenomenon contributes to GPT-4V’s relatively lower performance on metrics such as BLEU which primarily assesses word-match rates. One example is shown in Figure 5.6.

Clinical Efficacy on MIMIC-CXR Dataset: We assess the clinical efficacy of the GPT-4V model on the MIMIC-CXR dataset. We used the CheXbert tool <sup>1</sup> to label both the generated reports and the ground truth reports, obtaining results for 14 disease terms. We then calculated the accuracy, recall, and F1 score based on these generated labels. GPT-4V achieves a precision of 0.353, a recall of 0.365, and an F1 score of 0.330

<sup>1</sup><https://github.com/stanfordmlgroup/CheXbert>

in Table 5.2. Numerically, these results are not too bad compared to the two high-performing models METransformer [104] and R2GenGPT [105]. GPT-4V demonstrates competitive clinical efficacy. Notably, the gap between GPT-4V and other models in terms of clinical efficacy is relatively smaller than that observed in traditional NLP metrics, which primarily measure lexical overlap with ground truth reports. This suggests that, objectively, GPT-4V exhibits an impressive capability in radiology report generation, emphasizing its potential clinical accuracy and practicality.

**Table 5.2:** Evaluation of Clinical Efficacy on MIMIC-CXR dataset.

| Models              | Precision    | Recall       | F1           |
|---------------------|--------------|--------------|--------------|
| METransformer [104] | 0.364        | 0.309        | 0.334        |
| R2GenGPT [105]      | <b>0.392</b> | <b>0.387</b> | <b>0.389</b> |
| GPT-4V              | 0.353        | 0.365        | 0.330        |

### Medical Visual Question Answering

Table 5.3 shows a performance comparison between the GPT-4V model and SOTA VQA methods using the VQA-RAD dataset. The compared methods include StAn [34], BiAn [34], MAML [29], MEVF [75], MMQ [26], PubMedCLIP [28], MMBERT [49], and the Q2ATransformer [69]. Recall that we report accuracies for close-end questions and BLEU scores for open-end questions. From Table 5.3, GPT-4V achieves an accuracy of 61.4% on close-ended questions, which is notably lower than previously published results. In our experimental setup, we carefully categorize questions into open- and close-ended types and design prompts accordingly, aiming to guide GPT-4V’s response format. Despite this effort, we observe that GPT-4V occasionally produces overly verbose answers to close-ended questions that typically require succinct responses such as “yes” or “no.” This deviation from the expected format contributes to the observed drop in accuracy.

In addition, GPT-4V attains a BLEU score of only 0.116 for open-ended questions. This modest score stems from the model’s strong generative capacity, which often results in diverse but semantically valid responses. For instance, GPT-4V may express the same clinical concept

**Table 5.3:** Results on VQA-RAD benchmark

| Dataset             | Reference Methods | Fusion Method | Close-end |
|---------------------|-------------------|---------------|-----------|
| VQA-RAD             | StAn [34]         | SAN           | 57.2      |
|                     | BiAn [34]         | BAN           | 67.9      |
|                     | MAML [29]         | SAN           | 69.7      |
|                     |                   | BAN           | 72.4      |
|                     | MEVF [75]         | SAN           | 74.1      |
|                     |                   | BAN           | 75.1      |
|                     | MMQ [26]          | SAN           | 75.7      |
|                     |                   | BAN           | 75.8      |
|                     | PubMedCLIP [28]   | -             | 80        |
|                     | MMBERT [49]       | -             | 77.9      |
| Q2ATransformer [69] | -                 | 81.2          |           |
| GPT-4V [77]         | -                 | 61.40         |           |

using different phrasing, which negatively affects n-gram overlap-based metrics such as BLEU. These metrics, while widely adopted, are inherently limited in their ability to capture semantic equivalence and contextual relevance, especially in medical domains where variability in expression is common.

To address these limitations, we incorporate human evaluation as described in Section 5.3.4. Human reviewers provide a more nuanced and context-aware assessment of GPT-4V’s outputs, capturing aspects of clinical accuracy, coherence, and interpretability that automated metrics often miss. This dual evaluation approach enables a more comprehensive and reliable assessment of model performance, particularly in settings where linguistic variability does not necessarily imply reduced quality.

### Medical Visual Grounding

Table 5.4 reports a comparative performance of GPT-4V against various SOTA visual grounding methods using the MS-CXR dataset. We compare GPT-4V with a range of advanced methods in the field: BioViL [9], BioViL-T [8], RefTR [55], VGTR [27], SeqTR [128], TransVG [24], and MedRPG [15]. Each of these models represents a significant approach or

innovation in medical image analysis, making them suitable benchmarks for evaluating GPT-4V’s performance. By using the mIoU metric, we can quantitatively assess how well GPT-4V and other models perform in terms of accurately identifying and delineating relevant patterns within the medical images of the MS-CXR dataset. GPT-4V’s performance on the MS-CXR dataset yields a mIoU of 0.083, significantly lower than all published benchmarks. While GPT-4V demonstrates a level of comprehension in visual grounding, it struggles to accurately identify medical organs and pathological signs, leading to imprecise bounding box predictions. It should be noted that the recent SoM [115] model has shown substantial improvements in this area by segmenting and labeling images before grounding, enhancing performance on generic images. However, its effectiveness on medical images demanding finer details remains untested. Further research is necessary to assess its applicability to medical imaging.

**Table 5.4:** mIoU(%) results on MS-CXR benchmark.

| Methods | BioViL [9] | BioViL-T [8] | RefTR [55] | VGTR [27] | SeqTR [128] | TransVG [24] | MedRPG [15]  | GPT-4V [77] |
|---------|------------|--------------|------------|-----------|-------------|--------------|--------------|-------------|
| mIoU(%) | 22.9       | 24.3         | 50.11      | 53.58     | 56.63       | 58.91        | <b>59.37</b> | 8.33        |

### Human Evaluation Results

This section details our approach to evaluating and analysing human assessments of radiology reports and VQA tasks. We choose not to include human evaluation for the visual grounding task. The reason is that the output of visual grounding is typically a simple bounding box, rather than text. The quality of these bounding boxes can be straightforwardly assessed through visualization, making human evaluation less critical for this aspect. However, when it comes to the evaluation of text generation, visual inspection of the words alone is insufficient for determining quality. The key aspect of evaluating the text produced by a large-scale model like GPT-4V lies in understanding the semantic meaning of the generated content. Text quality encompasses more than just the correct words. Therefore, we incorporate human evaluation as a crucial part of our assessment process for text generation. Human evaluators can provide insights into how well the model captures the intended meaning,

context, and subtleties that traditional metrics may overlook. This approach allows for a more nuanced and comprehensive evaluation of the quality of text generated by GPT-4V.

### **Radiology Report Human Evaluation**

We randomly select 100 report pairs, including corresponding ground truth reports and the reports generated by GPT-4V. These pairs undergo grading by a trained radiologist, categorized into three tiers — high, medium, and low — based on varying degrees of human-perceived consistency. The scoring was carried out based on an integrated assessment of seven key criteria, each reflecting critical aspects of the quality and accuracy of the predicted reports. These criteria are:

- **Impression Consistency:** Evaluating the consistency of the impressions with the findings described.
- **Impression Organ:** Assessing the accuracy in identifying and describing the organ systems involved.
- **Description of Lesion:** Checking the precision and detail in the description of lesions.
- **Clinical History:** Considering the integration and relevance of the clinical history in the report.
- **Completeness:** Ensuring the report includes all necessary information and observations.
- **Grammar:** Reviewing the grammatical correctness of the report.
- **Medical Terminology:** Verifying the appropriate and accurate use of medical terminology.

Concurrently, we seek to examine the relationship between human ratings and conventional metrics. To achieve this, we calculate NLP scores, including BLEU, ROUGE, METEOR, and CIDEr, alongside clinical scores for each of the 100 report pairs. In the following analysis, we focus on BLEU and CIDEr due to the minimum variation in ROUGE and METEOR.

Visualization & Distribution Analysis: Among the 100 reports, the human rating identifies 10 high-quality, 35 medium-quality, and 55 low-quality reports. To assess and compare these findings against conventional evaluation metrics, we normalise NLP scores and organise them, along with the F1 clinical score, into three quality tiers. Specifically, the NLP scores undergo normalization, scaling them to the  $[0, 100]$  range using the formula  $\hat{s} = \frac{s - \text{lowest}}{\text{highest} - \text{lowest}} \times 100$ , where  $s$  is the original score, and  $\hat{s}$  is the scaled score. Subsequently, both the scaled NLP scores and the F1 score ( $\times 100$ ) are discretized into three quality ranges: low  $[0, 40)$ , medium  $[40, 70)$ , and high  $[70, 100]$ . The distributions of the 100 report samples are shown in Figure 5.7, corresponding to the evaluations by BLEU, CIDEr, F1, and human rating, respectively. Specifically, based on BLEU scores, there are 1 high-quality, 2 medium-quality, and 97 low-quality reports; based on CIDEr, there are 5 high-quality, 53 medium-quality, and 42 low-quality reports; and based on F1 score, there are 3 high-quality, 29 medium-quality, and 68 low-quality reports. Our findings indicate that traditional evaluation methods tend to yield lower scores compared to radiologist assessments, suggesting that human evaluation assigns higher scores to the reports generated by GPT-4V, implying better report quality than indicated by traditional scores.

Additionally, Figure 5.8 presents scatter graphs illustrating the relationship between the original BLEU/CIDEr/F1 scores and human ratings, respectively. The x-axis represents the original BLEU/CIDEr/F1 scores, while the y-axis depicts corresponding human scores for the same samples. Notably, a significant number of GPT-4V-generated reports with low BLEU scores ( $< 0.20$ ) receive "Medium" or "High" quality evaluations in human ratings, a trend also observed with CIDEr metrics. The human rating is better aligned with the F1 score reflecting clinical efficacy, e.g., there are fewer reports with low F1 scores ( $< 0.20$ ) rated as 'Medium' or 'High' by the human evaluator.

Correlation & Statistical Test: To further assess the correlation between the NLP scores and human ratings, we calculate Kendall's  $\tau$  coefficient (i.e., Kendall rank correlation coefficient), which measures the ordinal association between two quantities. Kendall's  $\tau$  is defined as

follows:

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{total number of pairs} \times (\text{total number of pairs} - 1)/2} \quad (5.1)$$

Concordant pairs occur when both the human evaluator and the NLP metric agree on the ranking of two reports.

Conversely, discordant pairs occur when there is a disagreement between the rankings by the human evaluator and the NLP metrics.

The  $\tau$  coefficient is within the range  $-1 \leq \tau \leq 1$ . If two random variables are independent, the expectation of  $\tau$  is zero. In addition, a statistical test with the null hypothesis of  $\tau = 0$  (no correlation) is conducted, and the p-value is reported. The p-value helps determine whether the observed agreement or disagreement in rankings between human ratings and NLP metrics is statistically significant or due to random variations.

As reported in Table 5, for the NLP scores BLEU, ROUGE, METEOR, and CIDEr, Kendall's  $\tau$  coefficients are small values around zero, suggesting the human rating and these NLP metrics are potentially independent. This is further supported by their corresponding large p-values, all exceeding 0.1, indicating a failure to reject the null hypothesis. Thus, the correlations between human ratings and NLP metrics are not statistically significant. However, the F1 score has a small p-value of 0.07 and a moderate Kendall's  $\tau$  of 0.144, showing a statistically significant rank correlation with the human rating. We also evaluated the RadCliQ score for GPT4v, which integrates multiple existing metrics. This score is obtained by linearly combining these metrics and adjusting the weights based on human-marked error scores to align more closely with human evaluations. The RadCliQ score's p-value is 0.015, showing a statistically significant rank correlation with the human rating. In short, the statistical tests support our observations that i) the quality of GPT-4V's generated reports could be much higher than what traditional evaluation methods indicate; and ii) GPT-4V's performance in clinical efficacy aligns with human evaluation, which is higher than NLP-based metrics. On the

other hand, it should also be noted that human evaluation captures more information than the clinical score of F1, as the latter is still limited in detecting the occurrence of clinically relevant keywords in the generated reports without comprehending the context.

An Example of Human Evaluation: To further bolster the credibility of our report generation results, we present an in-depth analysis conducted by a radiologist. This expert evaluation delves into the accuracy, relevance, and overall quality of the generated reports, providing valuable professional insights. The example in Figure 5.9 illustrates a notable disparity between BLEU scores and human evaluation. It juxtaposes a ground truth report with a report generated by GPT-4V, revealing that despite a low BLEU score of 0.069e-6, the report receives a high rating from a radiologist.

Upon expert examination, the radiologist affirms that the predicted report accurately identifies key bilateral lung consolidation areas, describes the patchy morphology of the lesions, and provides insightful pathological inferences. The changes in the mediastinal position of the heart are also consistent with the original report. Although the direct determination of cardiac hypertrophy is not made, a positive prompt is acknowledged, aligning with chest X-ray reporting principles. Other negative diagnosis descriptions are basically consistent with the ground truth report. Consequently, the generated report receives a high score based on these professional evaluations.

**Table 5.5:** Evaluation of P-Value and Kendall’s Tau

|               | Bleu-4 | ROUGE | METEOR | CIDEr | F1    | RadCliQ |
|---------------|--------|-------|--------|-------|-------|---------|
| P Value       | 0.688  | 0.430 | 0.462  | 0.503 | 0.07  | 0.015   |
| Kendall’s Tau | 0.032  | 0.063 | 0.059  | 0.053 | 0.144 | -0.193  |

### Visual Question Answering Human Evaluation

For this evaluation, we randomly select 100 open-end questions from VQA-RAD, together with ground truth and GPT-4V generated answers.

A trained radiologist assesses the correctness of GPT-4V's answers, identifying 57 answers as correct and 43 answers as incorrect. This contrasts sharply with the traditional accuracy metric used in the predominant classification-based VQA approaches, which detects only 9 correct answers, as shown in Figure 5.10. This disparity is rooted in the traditional classification-based VQA evaluation, which treats each predetermined answer as a class and calculates the classification accuracy. This evaluation lacks flexibility in handling the variability of answers produced by GPT-4V. Our human evaluation finds GPT-4V's answers more accurate than traditional scores, highlighting the richness of content in GPT-4V's responses not fully captured by conventional scoring methods. This underscores that the quality of GPT-4V's answers is not inferior.

Moreover, delving into the answers generated by GPT-4V, the radiologist finds GPT-4V not only matches but surpasses the accuracy of the ground truth. This superiority is attributed to the detailed and precise nature of GPT-4V's responses. Figure 5.11 provides two VQA examples produced by GPT-4V. In the right example, given the question "what organ system is visualized", the ground truth provided only mentions the chest, which is neither rigorous nor sufficient. In contrast, GPT-4V's response includes the skeleton and respiratory system, demonstrating improved accuracy and completeness. Moreover, the answers generated by GPT-4V exhibit a higher level of generalization and comprehensiveness. As shown in the left example, GPT-4V describes "mediport" as "A foreign body or medical device". Additionally, GPT-4V produces answers that are more professional and readable, presenting complete sentences with medical terminology and grammatical accuracy. These findings suggest that GPT-4V holds significant learning potential, inviting further exploration through additional research.

### **Case Study**

In this part, we demonstrate GPT-4V's performance by incorporating case studies across the three specific tasks, offering a more direct and intuitive view of its capabilities.

### Radiology Report Generation Cases

In zero-shot scenarios, through a series of tests on multiple chest X-ray images, it is observed that GPT-4V consistently generates reports with a focus on specific anatomical organs. This phenomenon can be observed in Figure 5.12. Notably, GPT-4V tends to present information in a specific sequence, covering the information on the lung, cardiomeastinal silhouette, bones, diaphragm, and soft tissues, in the majority of the generated reports. Examining the third case in Figure 5.12, GPT-4V demonstrates its proficiency to identify both normal and abnormal aspects, e.g., *"No pleural effusion or pneumothorax is present; Suggestive of a possible infectious or inflammatory process"*. These instances underscore GPT-4V's ability, even with zero-shot prompts, to generate relevant reports and identify anomalies.

In few-shot scenarios, our observation indicates that different prompts significantly influence the generated reports. GPT-4V's inclination to generate normal or abnormal findings varies based on the provided example reports. Figure 5.13 illustrates the response to a normal chest X-ray image, utilizing three distinct prompt settings to guide GPT-4V in generating corresponding reports. Interestingly, it is found that the report generated either from the normal-example prompt or the mixed-example prompt describes the image as normal, which is consistent with the ground truth. In contrast, the report from the abnormal-example prompt misidentifies anomalies in the image. Meanwhile, Figure 5.14 showcases an example of an abnormal chest X-ray image. It is observed that the report generated by the normal-example prompt misinterprets the lungs as normal. Our investigation emphasizes the substantial impact of the mixed-example prompt on GPT-4V's accuracy in determining the normality or abnormality of an image.

### Visual Question Answering Cases

We delve into specific cases of VQA illustrated in Figures 5.15, and 5.16. Figure 5.15 presents two examples of the close-end questions with a single-word answer of "yes" or "no". Effectively addressing such questions requires GPT-4V to accomplish two key tasks: firstly, discern that

the question is close-end, and secondly, comprehend the attributes of a close-end question. This proficiency is crucial because, despite the simplicity of these questions, GPT-4V tends to generate more extended responses, which will be deemed incorrect when assessing the accuracy of answers to close-end questions. Thus, GPT-4V faces the challenge of identifying when a question demands a concise one-word answer and adjusting its response generation accordingly. Properly tailoring its output to match the question type is essential for GPT-4V's accuracy in dealing with close-end questions.

Open-end questions demand more intricate responses beyond a simple "yes" or "no", providing flexibility in generating answers. However, for the VQA task, the challenge lies in predefined answer pools, making it hard for GPT-4V to generate an exact sentence that matches these predefined answers. In Figure 5.16, we present two examples of open-end questions. The first example showcases a correct, succinct response from GPT-4V, consisting of just one word. This showcases GPT-4V's effective handling of simple, direct open-end questions. The second example, however, depicts an incorrect answer by GPT-4V. The response provided by GPT-4V is a detailed explanation of the brain structures potentially affected by a lesion in the right frontal area, encompassing the prefrontal cortex, primary motor cortex, and possibly the premotor and precentral gyrus. This response appears more fitting for a theoretical question about brain anatomy, such as "What brain structures would be affected by a lesion in the right frontal area of the brain?" In this case, GPT-4V seems to have relied solely on medical knowledge without adequately considering the specific image in question. This suggests that GPT-4V might sometimes overlook the visual information in VQA tasks, especially when the question can be addressed through general medical knowledge. This insight underscores the challenge GPT-4V faces in generating responses for open-end VQA tasks.

### **Visual Grounding Cases**

In Figure 5.17 we provide two examples that shed light on GPT-4V's

performance in visual grounding tasks. Our examination suggests that while GPT-4V exhibits the capability to generate bounding boxes, its performance is suboptimal, especially in precisely locating objects within images. This limitation might stem from GPT-4V's challenges in processing and interpreting detailed image information, particularly in the context of medical images where a nuanced focus on fine-grid features is crucial for accurate visual grounding. Furthermore, we posit that GPT-4V's training, predominantly on common images, may contribute to its suboptimal performance with medical images. The model's limited exposure to diverse and labeled medical data might be a key factor in this inadequacy.

**VQA Prompt**

Input an X-ray image along with a question related to that image. Questions can fall into one of two categories: closed-ended or open-ended.

Closed-ended questions require a "Yes" or "No" response.  
Open-ended questions require an answer that may consist of a single word or a sentence, providing a more detailed explanation.

Based on the X-ray image and the accompanying question, determine whether the question is closed-ended or open-ended and provide an appropriate response.

Examples

Closed-ended Question  
question: Is subcutaneous air present in the right side of the neck?  
answer: Yes

question: Is there evidence of metastatic tissue on organs surrounding the liver?  
answer: no

Open-ended Question  
question: Describe the pathology visible in the image.  
answer: Cardiomegaly with pulmonary edema.

question: What is the vascular distribution of these infarcts?  
answer: all three vascular distributions

question: How large is the mass?  
answer: 5mm

question: What modality is used in this image?  
answer: CT

question: How many kidneys are visualizable in this image?  
answer: 2

Please directly provide the answer to the question without any additional explanations or information.

question: {{question}}

**Figure 5.4:** VQA Prompt Example. The content between the double braces is replaced with specific questions.

**VG Prompt**

You are a radiological AI assistant. I will provide a chest x-ray image and a descriptive text related to the image. please help to locate the region mentioned in the text on the image and output the coordinates for that region.

I will also provide you with the dimensions of the image. Please use this information to calculate the coordinates.

Coordinates should be presented as: [top left x, top left y, width, height].

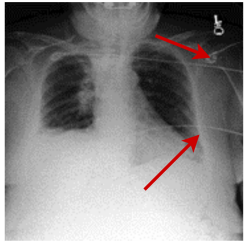
Please directly provide the coordinates without any additional explanations or information.

=====

Image Dimensions: width: **{{Width}}**, height: **{{Height}}**  
 The descriptive text is: **{{description}}**

**Figure 5.5:** VG Prompt Example. The content between double braces is replaced with specific image width, height, and description text related to the image.

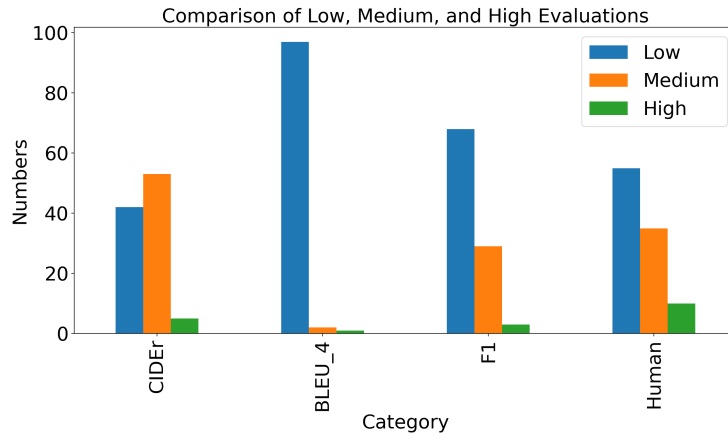
**Image:**



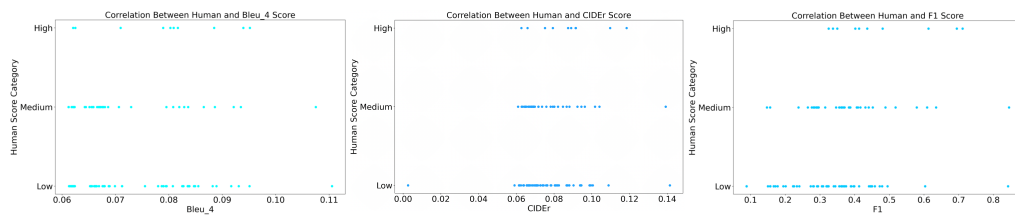
**Ground Truth:**  
 impression: Subtle left lower lobe opacity, which may represent atelectasis, but pneumonia cannot be excluded. PA and lateral radiographs could allow for better assessment of this opacity. Stable right pulmonary effusion. Findings: There is a right pleural effusion which is unchanged since prior exam. Again seen is a right hilar opacity consistent with fibrosis, better assessed on recent CT. A subtle left lower lobe opacity is seen, which may represent atelectasis, but pneumonia cannot be excluded. The lungs are otherwise clear. The cardiomeastinal silhouette is unchanged from prior exam. Visualized osseus structures are unremarkable.

**GPT-4V:**  
 Impression: Bilateral mid to lower lung field opacities, suggestive of potential consolidative process or infiltrates. **Presence of a medical device in the superior aspect of the chest.** Findings: The cardiomeastinal silhouette is within normal limits for contour. There are opacities observed in the bilateral mid to lower lung fields which may represent consolidation or infiltrates. The pulmonary vasculature appears within normal limits. An overlying medical device, likely a central venous catheter or similar device, is seen in the superior aspect of the chest. No evidence of pneumothorax. No visible pleural effusion. The osseous structures of the chest wall and visualized portions of the thoracic spine appear intact without any obvious deformity.

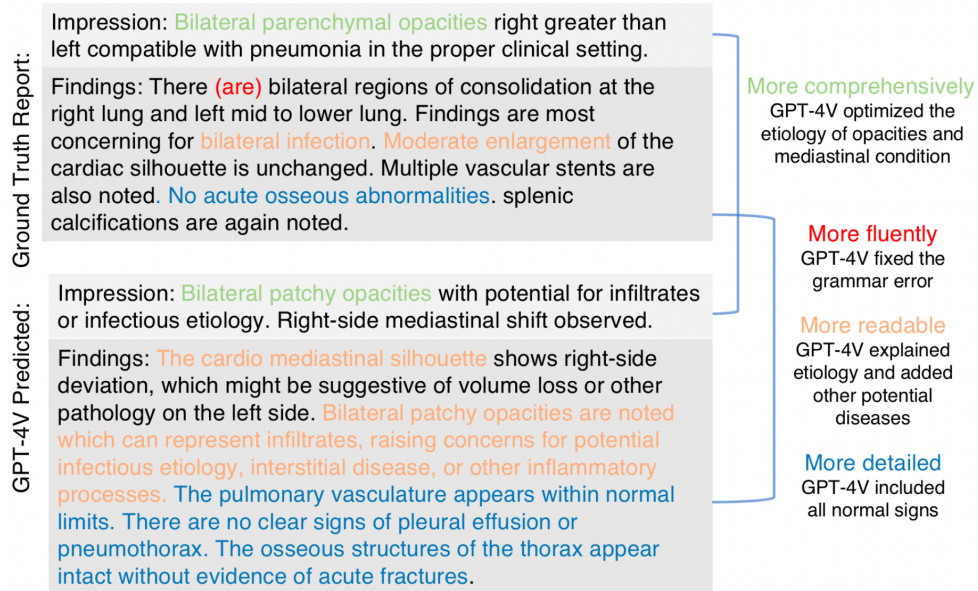
**Figure 5.6:** An illustration of GPT-4V's capability in generating medical reports using our designed few-shot mixed-example prompt. The ground truth does not mention a medical device but one is visibly present in the image, marked by red arrows. GPT-4V demonstrates the ability to recognize and describe the medical device in the generated report.



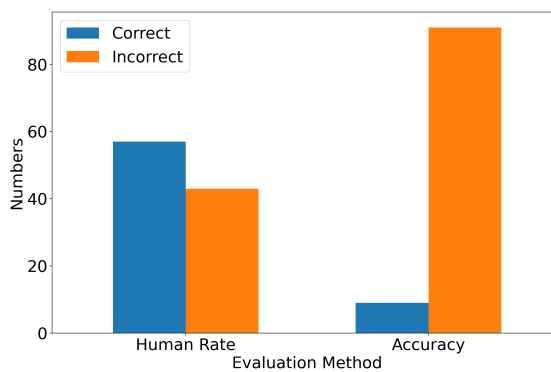
**Figure 5.7:** Distributions of scores based on human evaluation, BLEU, CIDEr, and F1, respectively.



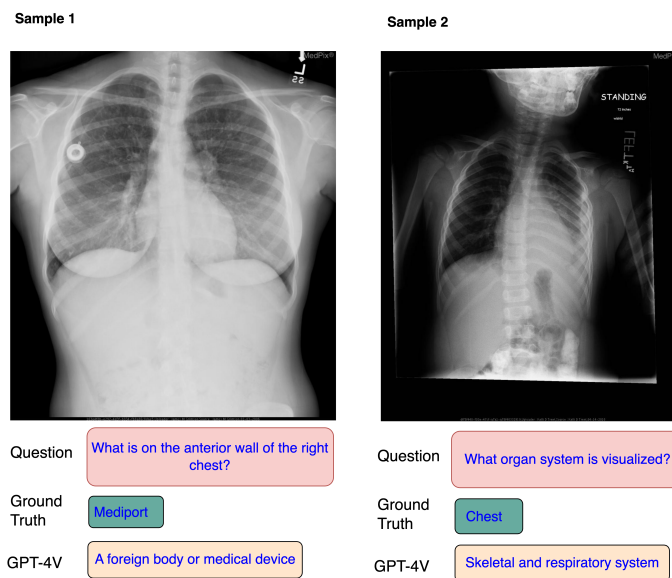
**Figure 5.8:** Scatter graphs show the relationship between BLEU/-CIDEr/F1 score and the human evaluation, respectively.



**Figure 5.9:** An example of a ground truth report and a GPT-4V generated report. Key medical information in the reports is highlighted using different colors.




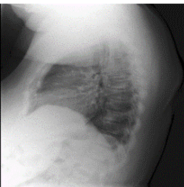

**Figure 5.10:** Human evaluation of the correctness of the answers generated by GPT-4V on a subset of VQA-RAD, in contrast to the evaluation based on the classification accuracy of the same dataset.



**Figure 5.11:** Two examples showcase GPT-4V's performance on open-end questions.

**Radiology Report Generation Case**

*You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. The report needs to contain both the impression and findings sections. You should only generate reports and no other unnecessary information.*

|   |  |
|---|--|
|    | <p><b>Ground Truth:</b><br/>           Impression: No acute intrathoracic process. Findings: The lungs are well expanded and clear. The cardiomeastinal silhouette, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. Sternal wires are intact.</p>   |
|   | <p><b>Ground Truth:</b><br/>           Impression: No acute intrathoracic process. Findings: The lungs are hypoinflated, accounting for vascular crowding but no evidence of focal opacities. Cardiomeastinal and hilar contours are unremarkable. The cardiac size is top normal. There is no pleural effusion or pneumothorax. A right subclavian stent is noted and unchanged from prior examination.</p>   |
|  | <p><b>Ground Truth:</b><br/>           Impression: Increased opacity projecting over the right hemi thorax likely due to increased right pleural effusion with overlying atelectasis, underlying infectious process not excluded. Findings: There relatively low lung volumes. There is increased opacity projecting over the right hemi thorax likely due to increased right pleural effusion with overlying atelectasis, underlying infectious process not excluded. Possible trace left pleural effusion. The cardiac silhouette is top-normal to mildly enlarged. Mediastinal contours are unremarkable. No pneumothorax is seen.</p>  |
|   | <p><b>GPT-4V:</b><br/>           Impression: Normal chest radiograph with no acute cardiopulmonary abnormalities. Findings: Lungs: Clear bilaterally. No focal opacity, consolidation, or mass identified. Cardiomeastinal silhouette: Normal size and contour. No mediastinal widening. Bones: Intact without any appreciable fractures or bony lesions. Diaphragm: Intact with no evidence of diaphragmatic hernia. Soft tissues: Unremarkable.</p> <p><b>GPT-4V:</b><br/>           Impression: Evidence of increased radiodensity in the mid and lower lung fields, which could suggest possible consolidation or infiltrates. Further clinical correlation is recommended. Findings: Lungs: Increased radiodensity in the mid and lower lung fields indicating possible consolidation or infiltrative pathology. Cardiomeastinal silhouette: Not entirely visualized, but no obvious abnormalities seen. Bones: Thoracic bony structures are intact with no signs of acute fractures. Diaphragm: Not explicitly mentioned in the report, but we could assume it's intact given no specified abnormalities. Soft tissues: The visible upper abdominal structures are unremarkable.</p> <p><b>GPT-4V:</b><br/>           Impression: Bilateral interstitial and alveolar opacities, predominantly in the upper and mid lung zones, suggestive of a possible infectious or inflammatory process. No overt pleural effusion or pneumothorax. Findings: The lungs are hyper-inflated with no focal consolidation. Bilateral interstitial and alveolar opacities are observed, more pronounced in the upper and mid lung zones. The cardiac silhouette is within normal limits for size and contour. No visible mediastinal or hilar lymphadenopathy. No evidence of pneumothorax or significant pleural effusion. The osseous structures of the chest are intact without visible fractures or lytic lesions.</p> |

**Figure 5.12:** R2Gen case using zero-shot prompt. GPT-4V can generate radiology reports without example reports and convey both normal and abnormal aspects. For better illustration, the key medical information in the reports is highlighted using different colors.

**Radiology Report Generation Case**

**Text-Prompt:**


*You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. You should only generate reports and no other unnecessary information. The format of the output report should be similar to the example:*

*Examples:*  
 =====

*Example 1*

*Example 2*

**Image:**



---

**Ground Truth:**  
 Impression: No acute intrathoracic process.  
 Findings: The lungs are well expanded and clear. The cardiomeastinal silhouette, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. Sternal wires are intact.


**Few-shot mixed examples prompt:**  
 Impression: Normal cardiopulmonary findings with no acute abnormalities. Findings: Frontal view of the chest demonstrates clear lung fields without focal consolidations, masses, or opacities. The cardiomeastinal silhouette and pulmonary vasculature are within normal limits. Both hemidiaphragms are intact with no evidence of elevation. There is no visible pneumothorax, pleural effusion, or significant lymphadenopathy. The bony thoracic structures appear intact without any evident pathological changes.

**Few-shot normal examples prompt:**  
 Impression: No acute cardiopulmonary abnormalities noted. Findings: The lungs are clear without any evidence of consolidation, pleural effusion, or pneumothorax. The cardiomeastinal silhouette is within normal limits. There is no evidence of pulmonary nodules or masses. No significant bony abnormalities are noted in the imaged portions of the chest. The imaged upper abdomen appears unremarkable.

**Few-shot abnormal examples prompt:**  
 Impression: Increased opacity in the left lower lung zone, consistent with consolidation or infiltrate, which may suggest pneumonia. Mediastinal and cardiac contours are within normal limits. Findings: PA view of the chest demonstrates an increased opacity in the left lower lung zone, which may be secondary to consolidation or infiltrate. The lungs are otherwise well aerated. No evidence of pneumothorax or pleural effusion. The mediastinal and cardiac contours appear within normal limits. No visible pulmonary nodules or masses. No evidence of mediastinal widening or adenopathy. The bony thorax and soft tissues appear unremarkable.



**Figure 5.13:** R2Gen normal case. Key medical information in the reports is highlighted using different colors. GPT-4V is more likely to generate reports containing abnormality descriptions when the prompt consists of only abnormal examples. The text in red corresponds to descriptions of abnormal conditions.

**Radiology Report Generation Case**

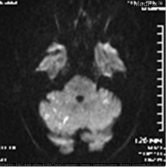

|  |  |
|--|--|
| <p><b>Text-Prompt:</b></p> <p><i>You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. You should only generate reports and no other unnecessary information. The format of the output report should be similar to the example:</i></p> <p><i>Examples:</i><br/>             =====</p> <p><i>Example 1</i></p> <p><i>Example 2</i></p>  | <p><b>Image:</b></p>  |
| <p><b>Ground Truth:</b></p> <p>Impression: <b>Increased opacity</b> projecting over the right hemi thorax likely due to increased right pleural effusion with overlying atelectasis, underlying infectious process not excluded. Findings: There relatively low lung volumes. <b>There is increased opacity projecting over the right hemi thorax</b> likely due to increased right pleural effusion with overlying atelectasis, underlying infectious process not excluded. Possible trace left pleural effusion. The cardiac silhouette is top-normal to mildly enlarged. Mediastinal contours are unremarkable. No pneumothorax is seen.</p>  |  |
| <p><b>Few-shot mixed examples prompt:</b></p> <p>Impression: <b>Bilateral interstitial opacities, potentially indicating interstitial lung disease or atypical infection.</b> Findings: <b>Frontal view of the chest shows bilateral interstitial opacities</b>, more prominent in the upper lobes. The cardiomeastinal silhouette is within normal limits. Pulmonary vasculature appears normal. Possible show pleural effusion. No evidence of pneumothorax. The diaphragms are at the appropriate levels, and bony thorax appears intact with no obvious fractures.</p>   |  |
| <p><b>Few-shot normal examples prompt:</b></p> <p>Impression: No acute cardiopulmonary abnormalities detected. Findings: <b>The lungs are clear without evidence of focal consolidation, pleural effusion, or pneumothorax.</b> The cardiomeastinal silhouette is within normal limits. No prominent hilar or mediastinal lymphadenopathy is seen. There is no evidence of pulmonary edema. The visualized bony structures appear intact without acute fractures or lesions. The imaged upper abdomen is unremarkable.</p>   |  |
| <p><b>Few-shot abnormal examples prompt:</b></p> <p>impression: <b>Bilateral patchy opacities observed</b>, possibly indicative of a pneumonia or interstitial lung disease. Elevated right hemidiaphragm. Findings: The provided chest X-ray demonstrates bilateral patchy opacities predominantly involving the upper and mid lung zones. The right hemidiaphragm appears elevated compared to the left, suggesting possible diaphragmatic dysfunction or underlying right basal atelectasis. The heart size appears within normal limits. No overt pleural effusion is noted. The cardiomeastinal silhouette is unremarkable without any significant lymphadenopathy. The osseous structures are intact with no evident bony pathology.</p> |  |

**Figure 5.14:** R2Gen abnormal case. Key medical information in the reports is highlighted using different colors. GPT-4V is more likely to generate normal reports when the prompt consists of only normal examples. The text in red corresponds to descriptions of normal conditions.

**Visual Question Answering Cases (Close-ended Questions)**


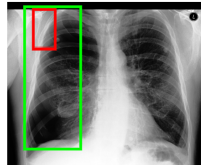

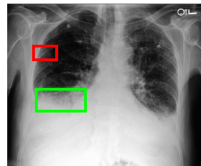
|  |  |
|--|--|
| <p><b>Text-Prompt:</b><br/> <i>Input an X-ray image along with a question..... (VQA Prompt)</i><br/>             =====<br/> <b>question: Is this patient in a supine position?</b></p>   | <p><b>Text-Prompt:</b><br/> <i>Input an X-ray image along with a question..... (VQA Prompt)</i><br/>             =====<br/> <b>question: Is the lesion a solid or cystic lesion?</b></p> |
|  <p><b>Ground Truth:</b><br/>Yes</p> <p><b>GPT-4V:</b><br/>Closed-ended Question<br/>answer: Yes.</p> |  <p><b>Ground Truth:</b><br/>No</p> <p><b>GPT-4V:</b><br/>Closed-ended Question<br/>answer: No.</p>   |

**Figure 5.15:** VQA case examples for close-ended questions. By few-shot prompts, GPT-4V could discern the question type and generate correct "yes" or "no" answers.

| Visual Question Answering Cases<br>(Close-end Questions)   |  |
|--|--|
| <p><b>Text-Prompt:</b><br/>                     Input an X-ray image along with a question..... (VQA Prompt)<br/>                     =====<br/>                     question: What organ system is shown in the above image?</p>   | <p><b>Ground Truth:</b><br/>                     Brain</p> <p><b>GPT-4V:</b><br/>                     Open-ended Question<br/>                     answer: Brain.</p>  |
| <p><b>Text-Prompt:</b><br/>                     Input an X-ray image along with a question..... (VQA Prompt)<br/>                     =====<br/>                     question: What brain structures would be affected by the lesion in the right frontal area of the brain?</p>  | <p><b>Ground Truth:</b><br/>                     basal ganglia (caudate and putamen)</p> <p><b>GPT-4V:</b><br/>                     Open-ended Question<br/>                     answer: The lesion in the right frontal area of the brain would affect the right frontal lobe structures including the prefrontal cortex, primary motor cortex, and possibly the premotor and precentral gyrus.</p> |

**Figure 5.16:** VQA case examples for open-end questions. By using few-shot prompts, GPT-4V has the capacity to generate correct answers (left). However, it may sometimes overlook visual information and generate answers solely based on general medical knowledge (right).

Visual Grounding Cases

|  |   |
|--|---|
| <p><b>Text-Prompt:</b><br/> <i>You are a radiological AI assistant.....(VG Prompt)</i><br/>         =====<br/> <b>Image Dimensions:</b> {width: 1767, height: 2021}<br/> <b>The descriptive text is:</b> <i>Large right-sided pneumothora</i></p> <p><b>GPT-4V:</b><br/>         [top left x: 400, top left y: 100, width: 350, height: 600]</p> <p><b>Ground Truth:</b><br/>         [265, 90, 880, 2198]</p>                 | <p><b>Image Prompt:</b></p>  <p><b>Bounding Box:</b></p>     |
|  |   |
| <p><b>Text-Prompt:</b><br/> <i>You are a radiological AI assistant.....(VG Prompt)</i><br/>         =====<br/> <b>Image Dimensions:</b> {width: 2539, height: 3050}<br/> <b>The descriptive text is:</b> <i>Bilateral multifocal areas of consolidation</i></p> <p><b>GPT-4V:</b><br/>         [top left x: 130, top left y: 350, width: 650, height: 300]</p> <p><b>Ground Truth:</b><br/>         [1510, 1268, 755, 796]</p> | <p><b>Image Prompt:</b></p>  <p><b>Bounding Box:</b></p>  |

**Figure 5.17:** Visual Grounding Prompt case. The bounding boxes in red are predicted by GPT-4V, while the ground truth bounding boxes are in green.



## Chapter 6

# MRScore: Evaluating Radiology Report Generation with LLM-based Reward System

Building upon the conclusions drawn in Chapter 6, this chapter delves deeper into the limitations of conventional NLG evaluation metrics in the context of radiology report generation. Despite recent advances in automated radiology report generation, evaluating the clinical and semantic quality of generated reports remains an open challenge. Widely used natural language generation (NLG) metrics such as BLEU, ROUGE, and METEOR often fail to capture the nuanced clinical correctness and semantic coherence that are crucial in medical settings. Our observations and analysis presented in earlier chapters further substantiate this inadequacy. To address these limitations, we propose **MRScore**, a novel automatic evaluation metric specifically designed for radiology report generation by harnessing the semantic capabilities of large language models (LLMs). Unlike traditional surface-level metrics, MRScore is designed to reflect human-level assessment more accurately, emphasizing clinical relevance and factual correctness. The development of MRScore is guided by collaboration with board-certified radiologists and consists of two key components:

- **LLM-guided data generation:** We employ GPT to synthesize a large-scale dataset comprising radiology reports of varying quality,

simulating both high-quality (accepted) and low-quality (rejected) samples.

- **Pairwise preference training:** These accepted-rejected pairs are used to train a reward model that assigns scalar scores MRScore, reflecting the report’s quality from a clinical and linguistic perspective.

Experimental results demonstrate that MRScore achieves a significantly higher correlation with expert radiologist judgments compared to traditional metrics. Moreover, MRScore proves more effective for model selection, providing a robust and scalable solution for evaluating radiology report generation models in real-world clinical applications.

## 6.1 Introduction

Automated evaluation of text generation systems, such as those used in radiology report generation, typically involves comparing generated outputs against reference reports to assess semantic accuracy. However, widely adopted natural language generation (NLG) metrics such as BLEU [82] primarily quantify n-gram overlaps, often neglecting the lexical and structural diversity that is essential for preserving semantic meaning. Recent studies [56] have highlighted several critical limitations of these n-gram-based metrics.

First, such metrics often fail to accommodate paraphrasing due to their rigid reliance on surface-level pattern matching. For instance, BLEU and METEOR [88] may penalize semantically equivalent expressions that differ in phrasing, thereby underestimating valid report quality. To address this, metrics like BERTScore [124] employ contextualized token embeddings to better capture paraphrastic similarity.

Second, conventional NLG metrics struggle to capture clinically important diagnostic content in radiology reports. Consequently, clinical-specific evaluation tools, such as CheXbert F1 [92] or RadGraph F1 [41], have been introduced to supplement traditional metrics with domain relevance.

Third, and most critically, despite ongoing efforts to refine evaluation, existing metrics remain misaligned with human judgment [56]. To bridge this gap, RadCliQ [120] was recently proposed, combining BLEU, BERTScore, CheXbert, and RadGraph F1 via a linear regression trained on a limited number of human-annotated samples. While RadCliQ achieves improved alignment, it still suffers from limitations: it relies heavily on costly human-labeled data and places disproportionate weight on clinical correctness while underemphasizing linguistic fidelity.

To advance the field, we propose **MRScore**—a new metric specifically designed for evaluating automatically generated radiology reports. Developed in collaboration with professional radiologists, MRScore is grounded in a structured framework that explicitly models the expert rules and priorities used in clinical report assessment. Our method capitalizes on the semantic capabilities of large language models (LLMs) to generate training samples that emulate human evaluations, thereby enabling the construction of a large, high-quality dataset without requiring extensive manual annotation.

Compared with existing evaluation metrics, MRScore offers several key advantages:

- It significantly improves alignment with human evaluations, delivering more accurate and clinically meaningful quality assessments.
- It is scalable and low-cost, thanks to its reliance on GPT-generated pseudo-evaluations rather than manual labels<sup>1</sup>.
- It supports transparent and flexible evaluation criteria, allowing for easy adaptation to new clinical domains or reporting formats.

To operationalize our framework, we first assess the capacity of GPT-4 [77] to generate human-like evaluations using a carefully designed prompt template. For each of 1,000 ground truth reports, GPT-4 produces three reports of varying quality (low, medium, high) across seven clinical and linguistic dimensions, resulting in 3,000 synthetic samples. These

---

<sup>1</sup>Our framework does not require human-labeled training data, but uses one or two human evaluation examples in prompts to guide GPT’s output generation.

samples are then paired into accepted–rejected report pairs, where the accepted report is semantically and clinically superior by a clear margin. A reward model built on the Mistral-7B-Instruct [42] backbone is trained on these pairs to predict MRScore as a scalar reward.

To validate MRScore, we apply it to score 100 GPT-4V-generated radiology reports and compare the results against conventional metrics. Correlation analyses show that MRScore aligns more closely with human assessments than existing metrics, demonstrating its efficacy as a next-generation evaluation tool.

**Our main contributions are as follows:**

- (1) We demonstrate that GPT-4 can generate high-fidelity, human-like evaluations for radiology reports. Leveraging this capability, we propose a scalable method to construct large training datasets without manual annotation.
- (2) We introduce a structured evaluation framework that distills radiologists’ assessment criteria into seven key dimensions, improving the interpretability and reliability of automatic scoring.
- (3) We develop **MRScore**, a reward model trained using paired quality-controlled report samples. MRScore achieves significantly higher correlation with expert judgments than existing metrics, validating its effectiveness for radiology report evaluation.

## 6.2 Related Work

Radiology report evaluation metrics can be broadly classified into two types: linguistic metrics and clinical metrics.

**Linguistic metrics**—including BLEU [82], ROUGE-L [61], METEOR [6], CIDEr [100], and BERTScore [123]—quantify the textual similarity between generated and reference reports. BLEU measures the precision of n-grams (up to 4), using the geometric mean of n-gram precisions along with a brevity penalty, making it sensitive to surface-level word overlap. ROUGE-L focuses on recall by identifying the longest common

subsequence (LCS) between the candidate and reference texts, reflecting sentence-level structural similarity. METEOR enhances evaluation by accounting for synonyms and stemming, computing the harmonic mean of precision and recall to capture semantic overlap. CIDEr, which weights n-gram similarity by TF-IDF, emphasizes consensus with reference reports and is particularly suited for descriptive text. In contrast to surface-matching approaches, BERTScore employs contextual embeddings from BERT to compare token-level semantic similarity, providing more nuanced precision, recall, and F1 scores.

While these linguistic metrics are valuable for assessing fluency and textual alignment, they often fail to capture the correctness of medical content, especially in cases where lexical variations exist but the clinical meaning is preserved.

**Clinical metrics** are designed to directly evaluate medical correctness. CheXpert Score [40] extracts binary clinical labels (e.g., pneumonia, cardiomegaly) from both the generated and ground-truth reports, and computes F1 scores to assess agreement at the finding level. RadGraph Score [41] goes further by extracting structured entities and relations from reports and computing similarity using BERT embeddings, thereby evaluating the factual consistency of clinical concepts and their relationships.

To provide a more comprehensive assessment, **composite metrics** such as RadCliQ [121] aggregate multiple automated evaluation scores, including both linguistic and clinical aspects. RadCliQ is trained to correlate with radiologist judgments, thereby offering a unified metric that approximates expert-level evaluation of report quality.

These metrics, while widely adopted, still face limitations in capturing subtle clinical nuances or long-range dependencies in medical narratives. As such, ongoing research continues to explore more semantically aware and clinically grounded evaluation strategies.

## 6.3 Method

In the domain of automated radiology report generation, evaluating the quality of generated reports is typically achieved using a metric function  $f(x, \hat{x})$ , where  $x$  denotes the generated report and  $\hat{x}$  represents the corresponding reference report. Conventional metrics—including BLEU [82], ROUGE [61], METEOR [6], and CIDEr [100]—are widely used for this purpose. However, these metrics primarily measure surface-level n-gram overlap, often failing to capture deeper semantic equivalence and clinical correctness, which are essential for accurate medical report evaluation.

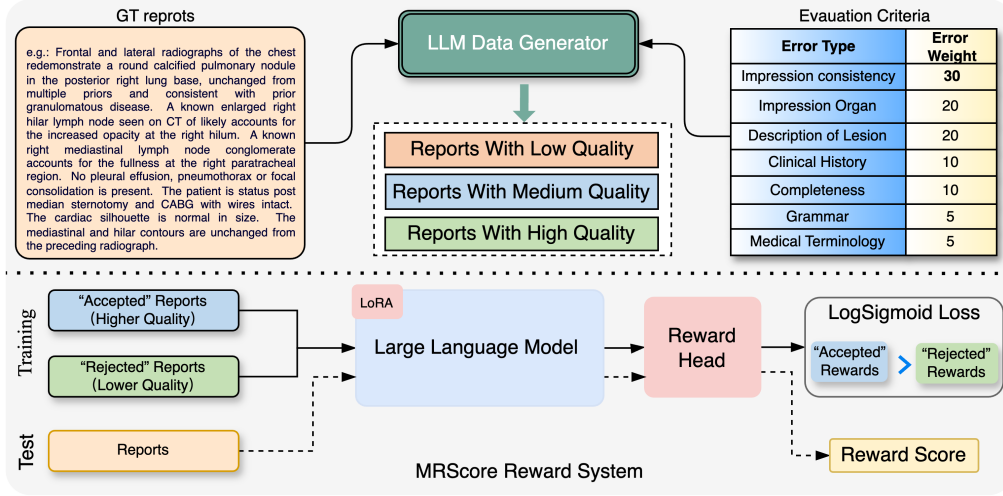
To overcome these limitations, we introduce a novel evaluation framework aimed at learning a semantically and clinically informed reward function, which we refer to as **MRScore**. This framework is designed to evaluate reports with human-aligned precision, without requiring manually annotated training data.

Figure 6.1 provides an overview of our approach. The framework consists of three main components: (1) the construction of a scoring dataset using GPT-generated reports across various quality levels; (2) the formation of preference-based training pairs to reflect relative report quality; and (3) the training of a reward model to output scalar scores that reflect alignment with expert judgment.

The subsequent sections detail each component of this framework.

### 6.3.1 Scoring Criteria Design and Scoring Dataset Generation

**Report Evaluation Criteria** We collaborated closely with radiologists to develop a reliable scoring system based on seven fundamental items from their expertise and literature review, outlined in Table 6.1. These items comprehensively cover both clinical findings and linguistic concerns, encompassing “impression consistency”, “impression organs”, “description of lesions”, “clinical history”, “completeness”, “grammar”, and “medical terminology”. A detailed explanation of each item is provided in Table 6.1. Our scoring system is error-based. Each item corresponds to an



**Figure 6.1:** The upper section depicts the process of generating synthetic training data using GPT-based report sampling across different quality levels. The lower section illustrates the reward model training procedure via LoRA adaptation. Solid arrows denote steps executed during training, while dashed arrows indicate the inference phase.

error type and is assigned a varied weight, as shown in Table 6.1. The total score is calculated as follows, within the range of  $[0, 100]$ :

$$\text{Total\_score} = 100 - \sum_{i=1}^7 S_i \times W_i. \quad (6.1)$$

Here  $S_i$  represents the individual error score for the  $i$ -th error item, with a value of 1 if an error occurs and 0 if not.  $W_i$  denotes the weight for the  $i$ -th error type, with specific values given in Table 6.1.

**GPT-4’s Evaluation** Based on our predefined scoring framework, we designed prompts that reflect each evaluation criterion, enabling GPT-4 to perform radiology report assessments in a human-like manner. To examine the alignment between GPT-4 and expert evaluations, we asked both a certified radiologist and GPT-4 to independently evaluate 100 reports generated by models with varying quality against their corresponding reference reports randomly drawn from the MIMIC-CXR dataset. Table 6.2 presents the Kendall’s Tau coefficients, which quantify the ordinal agreement between the two evaluators, along with the associated p-values from statistical significance testing. Remarkably, GPT-4’s scores

**Table 6.1:** Table for error types and design detail.

| Error Type (Weight)         | Explanation  |
|-----------------------------|--|
| Impression consistency (30) | Verify the presence of an ‘impression’ section in the report first; its absence deems the report of inferior quality, as it contains crucial diagnostic details from the X-ray. A clear and accurate ‘impression’ is vital for radiologists to provide exceptional care. Crafting it demands substantial effort, as it involves more than just summarizing findings, with inaccuracies potentially diminishing the report’s quality[33].   |
| Impression Organ (20)       | Evaluate the precision of the impression section where present. A high-quality report should detail how surrounding organs are affected, highlighting the need for precise organ descriptions, a practice underscored by literature that recommends using subheadings for each organ and system in the findings[30].   |
| Description of Lesion (20)  | Review the report to ensure the anatomical organ related to the lesion is correctly identified. Check that it correctly describes the lesion’s location, size, and opacity, as well as cardiovascular size and bone integrity, as per the ground truth. Descriptions of abnormal X-rays should include the lesion’s precise anatomical location, distribution pattern, edges, shape, internal structure, and size, along with the quantity of lesions, their density, and the affected surroundings. Use standard terms for the lesion’s location, describe its boundaries clearly, and measure the size of all significant lesions. For multiple lesions, record the size of the largest one, and describe the quantity and density accurately. Note any additional abnormalities within the scan area, and suggest further examination if the findings are inconclusive. |
| Clinical History (10)       | Ensure that the predicted report accurately mentions any operation history, treatment, and family history as outlined in the ground truth. Providing clinical history is essential for justifying radiological exams and allows radiologists to fine-tune their analysis to address clinical queries, ultimately leading to diagnoses that integrate patient history and imaging findings[87].   |
| Completeness (10)           | Verify the completeness of the predicted report. Completeness is as essential as accuracy in radiology reports, reflecting the radiologist’s expertise in providing detailed yet concise information vital for comprehensive patient care.   |
| Grammar (5)                 | Ensure the predicted report is free from spelling errors and is clearly articulated. Proper grammar, accurate spelling, precise language, and coherent structure are pivotal in radiology reports for ensuring clarity[109, 79]. Since these reports significantly influence professional evaluations and patient care, meticulous proofreading is crucial to prevent any misinterpretation[32].   |
| Medical Terminology (5)     | Verify the correct usage of medical terminology in the predicted report. This standardized language is crucial for clear communication among healthcare professionals, particularly for interpreting imaging reports and facilitating accurate diagnoses, treatment planning, and collaboration[73].   |

**Table 6.2:** Verification of GPT-4’s Capacity in Generating Human-like Ratings

| Metrics     | Kendall’s Tau ↑ (P value ↓) |
|-------------|-----------------------------|
| GPT-4       | 0.531 (5.98e-11)            |
| BERTScore   | 0.092 (2.04e-01)            |
| RadGraph F1 | 0.060 (4.32e-01)            |
| BLEU-4      | 0.048 (5.07e-01)            |
| ROUGE       | 0.120 (9.92e-02)            |
| METEOR      | 0.082 (2.58e-01)            |
| CIDEr       | 0.032 (6.54e-01)            |

show a strong positive correlation with the radiologist’s judgments, yielding a Kendall’s Tau value of 0.531. The extremely low p-value of 5.98e-11 underscores the statistical reliability of this result. Furthermore, GPT-4’s evaluations demonstrate markedly better alignment with human ratings than conventional NLG metrics and clinically oriented evaluation scores.

**Scoring Dataset Generation** Having verified GPT-4’s capability to deliver evaluations comparable to those of human experts, we then employed it to construct our scoring dataset. Leveraging the GPT-4 API, we synthesized a total of 3000 reports of varying quality levels, each derived from a randomly chosen set of 1000 reference reports from the MIMIC-CXR dataset. For every reference report, GPT-4 was prompted to produce three distinct versions corresponding to different quality ranges:

low (0–40), medium (40–70), and high (70–100). This stratified generation strategy was adopted to ensure a well-balanced distribution of scores across the entire range. After filtering out defective or unusable samples, we curated a final dataset of 2994 reports, each annotated with an associated evaluation score provided by GPT-4.

### 6.3.2 LLM-based Reward Model

Extending our prior findings, we introduce **MRScore**, a novel evaluation metric developed within our tailored evaluation pipeline. MRScore operates as a reward model based on large language models (LLMs), trained using the Reinforcement Learning with Human Feedback (RLHF) paradigm [93]. Specifically, we fine-tune a pretrained LLM, such as Mistral [42], to mirror human judgment by leveraging training pairs of <accepted, rejected> reports. Each pair originates from the same reference report, with the accepted version scoring higher under GPT-4’s evaluation than its rejected counterpart. The reward model is trained to preferentially assign higher scores to accepted samples. At inference time, the model outputs a predicted reward that serves as the MRScore, providing an automated and human-aligned assessment of report quality. The detailed structure of our framework is described below.

**Training Pair Construction** The training data pairs were derived from our GPT-4-generated scoring dataset, which is based on the error-aware evaluation scheme we established. Owing to the strong alignment between GPT-4 scores and human assessments, this dataset effectively approximates human preferences when ranking radiology reports. Each training instance is composed of a preferred report  $y_w$  and a less favored report  $y_l$ , both generated from the same reference report  $x$ , where  $y_w$  receives a higher evaluation score from GPT-4 than  $y_l$ . Using these instances, we construct two corresponding prompts: an ‘accepted’ prompt  $p(x, y_w)$  and a ‘rejected’ prompt  $p(x, y_l)$ , where  $p(\cdot)$  denotes the prompt formatting operation used to instruct the model during training.

**Backbone Model** Our reward model is built upon the Mistral-7B-instruct backbone [42], which is fine-tuned using Low-Rank Adaptation (LoRA)

[36], a widely adopted Parameter-Efficient Fine-Tuning (PEFT) technique [35]. This strategy allows us to align the model outputs with human-like judgments efficiently. We chose Mistral-7B-instruct due to its robust capabilities in language understanding and comprehension, which proved superior to other alternatives in our evaluations, making it a well-balanced choice between performance and computational cost.

**Training Objective** The purpose of the reward model is to replicate human preferences through accurate prediction. This is achieved by formulating a reward function that reflects the relative quality of candidate reports. For MRScore, this reward is derived from comparative rankings among report pairs, enabling the model to infer which report is more likely to be favored by humans. The detailed formulation of the loss function guiding this objective is described in Section 6.3.3.

**Training Procedure** During the fine-tuning stage, the pretrained LLM backbone is adapted to the specific task of reward estimation, enabling its outputs to better reflect human judgment. For MRScore, we fine-tuned the Mistral-7B-instruct model on our curated training pairs, training it to distinguish between high- and low-quality reports while considering the quality gap:  $margin = score_{accept} - score_{reject}$ . A larger margin denotes a greater disparity in report quality, whereas a smaller margin reflects more subtle differences.

To support this functionality, we append a reward head to the Mistral backbone—a linear layer that projects the LLM’s hidden representations into a scalar output. A sigmoid activation is then applied to normalize this output to the  $[0, 1]$  range, where a score of 1 indicates superior quality and 0 denotes poor quality. This scalar value is subsequently used to rate generated reports, and serves as our final MRScore metric.

The complete architecture and training pipeline are depicted in the lower section of Fig. 6.1.

### 6.3.3 Loss Function

The loss function guiding the training of our reward model is defined in Equation 6.2. In this formulation,  $\gamma_\theta(\cdot)$  represents the learnable reward function parameterized by  $\theta$ . The variable  $x$  denotes the reference (ground truth) report,  $y_w$  corresponds to the higher-quality (accepted) report, and  $y_l$  indicates the lower-quality (rejected) report. The margin  $m$  reflects the required reward gap between better and worse predictions.  $\log(\cdot)$  denotes the natural logarithm applied to the logistic function.  $D$  refers to the dataset of training samples, and  $K$  is the number of report samples in a training batch.

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} \sum_{(x, y_w, y_l) \sim D} [\log(\sigma(\gamma_\theta(x, y_w) - \gamma_\theta(x, y_l) - m))] \quad (6.2)$$

This loss encourages the model to assign significantly higher rewards to superior reports than to inferior ones, with a margin  $m$  reflecting the expected difference in quality. When the margin requirement is not met, the model is penalized, thereby promoting reward outputs that better mirror human preferences as captured by our GPT-4-annotated scoring dataset.

## 6.4 Experiments and Result

### 6.4.1 Evaluation Dataset

We evaluated the reliability of MRScore by analyzing how well its outputs align with expert assessments from radiologists, focusing on the correlation between its predicted scores and human judgment. To this end, we constructed a dedicated evaluation set composed of 100 radiology reports generated by GPT-4V [77], a state-of-the-art multimodal language model recognized for its seamless integration of image inputs and strong performance in image-grounded tasks. Prior studies have validated GPT-4V’s effectiveness in medical image understanding and its ability to produce clinically relevant radiology reports [56], making it a suitable candidate for generating a representative evaluation dataset.

Specifically, we selected 100 chest X-ray (CXR) images at random from the MIMIC-CXR dataset and used them as direct inputs to GPT-4V to generate corresponding reports.<sup>2</sup> A board-certified radiologist subsequently reviewed each of these generated reports, comparing them to their respective ground truth reports from MIMIC-CXR. Based on clinical accuracy and report quality, the radiologist assigned one of three quality levels—high, medium, or low—to each sample, corresponding to numerical scores of 90, 60, and 30, respectively.

## 6.4.2 Experiment Result

Table 6.3 presents the correlation between human assessments and various evaluation metrics, including traditional NLG metrics such as BLEU-4, METEOR, and CIDEr, as well as clinically oriented scores like BERTScore and RadGraph F1. Additionally, we include a comparison with RadCliQ-based scores, which are derived via regression from human-labeled error annotations.

The results clearly highlight the strong concordance of MRScore with expert evaluations, achieving the highest Kendall’s Tau coefficient (0.178) and Spearman correlation (0.221) among all metrics evaluated. These values notably outperform those of standard NLG metrics and clinically relevant benchmarks. While conventional NLG metrics exhibit weak or statistically insignificant associations with human ratings, clinical metrics such as BERTScore and RadGraph F1 demonstrate significant correlation, as evidenced by their lower p-values. This supports our metric’s broader evaluation framework, which captures both textual fidelity and medical accuracy, moving beyond strict token-level overlap to more meaningful clinical alignment.

It is also important to highlight the performance of RadCliQ-based scores, which combine multiple metrics, including BLEU, BERTScore, ChestXbert, and RadGraph F1. Despite being trained using a different set

---

<sup>2</sup>It is important to distinguish that these 100 reports were generated directly from CXR images using GPT-4V, and are distinct from the 100 reports referenced in Section 7.3.1, which were used to assess the correlation between GPT-4 evaluations and human ratings.

of human-labeled error scores, RadCliQ metrics still display significant alignment with radiologist assessments in our evaluation. This may be attributed to a shared focus on clinically important elements, reinforcing the value of integrating domain-specific insights into evaluation systems.

**Table 6.3:** Assessment of Human Correlations on Evaluation Dataset

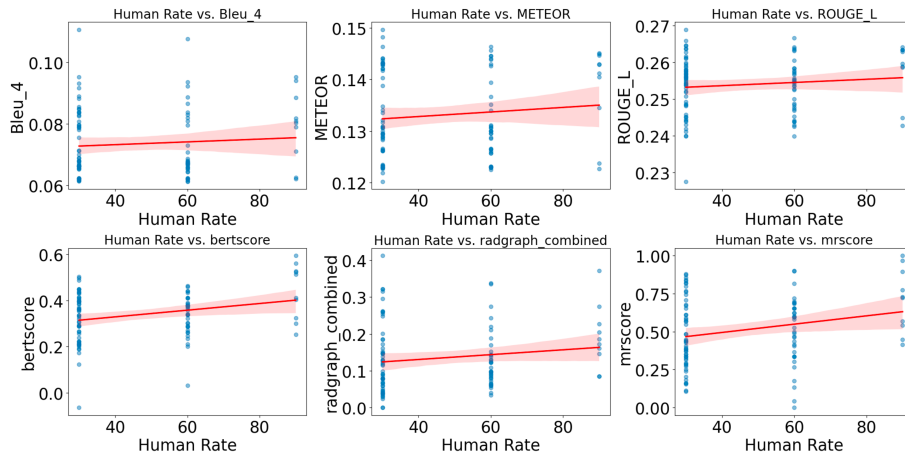
|                 | BLEU-4 | ROUGE | METEOR | CIDEr | BERTScore | RadGraph F1 | RadCliQ-V0 | RadCliQ-V1 | <b>MRScore</b> |
|-----------------|--------|-------|--------|-------|-----------|-------------|------------|------------|----------------|
| P Value ↓       | 0.688  | 0.429 | 0.463  | 0.503 | 0.0446    | 0.071       | 0.018      | 0.016      | <b>0.002</b>   |
| Kendall's Tau ↑ | 0.032  | 0.063 | 0.059  | 0.053 | 0.159     | 0.144       | -0.189     | -0.193     | <b>0.250</b>   |
| P Value ↓       | 0.677  | 0.484 | 0.460  | 0.422 | 0.045     | 0.080       | 0.0175     | 0.0015     | <b>0.002</b>   |
| Spearman ↑      | 0.042  | 0.071 | 0.074  | 0.081 | 0.200     | 0.176       | -0.236     | -0.241     | <b>0.304</b>   |

The scatter plots illustrated in Fig. 6.2 provide a visual comparison between various scoring metrics and expert human ratings for radiology reports. Each subplot corresponds to one specific metric, where individual points represent the scores assigned to reports plotted against the human-annotated quality levels. Trend lines with surrounding confidence bands reveal the strength and reliability of correlations. Among all evaluated metrics, MRScore exhibits the strongest positive correlation with expert scores, indicating its high consistency with radiologist assessments.

In contrast, conventional NLG metrics such as BLEU-4, METEOR, and ROUGE demonstrate weaker and more scattered relationships with human evaluations, reflecting inconsistencies in their ability to capture semantic and clinical accuracy. BERTScore, while positively correlated, does not achieve the same alignment as MRScore. Likewise, RadGraph F1 shows a positive association but still falls short of MRScore in terms of correlation strength.

Overall, the results suggest that MRScore provides a more faithful reflection of human judgment compared to existing metrics, reinforcing its potential as a reliable evaluation tool for radiology report generation.

Table 6.4 shows the performance of our reward model using various LLMs, including Phi-2 [59], Gemma-2b-it [31], and Gemma-7b-it [31], alongside our chosen Mistral-7b [42] as the backbone models. Once again, both Kendall and Spearman correlation coefficients are utilized for assessment. Among these models, Mistral-7b, with 6.8M trainable parameters, demonstrates the highest consistency with human ratings. It achieves a



**Figure 6.2:** Correlations between varied evaluation metrics and radiologist scores

**Table 6.4:** Human Correlations of MRScore Using Different LLM Backbones

| Model            | Trainable params (%) | Kendall $\uparrow$ (P value $\downarrow$ ) | Spearman $\uparrow$ (P value $\downarrow$ ) |
|------------------|----------------------|--|---|
| Phi-2 [59]       | 5.2M (0.197)         | 0.144 (0.069)                              | 0.178 (0.075)                               |
| Gemma-2b-it [31] | 1.8M (0.073)         | 0.135 (0.091)                              | 0.169 (0.092)                               |
| Gemma-7b-it [31] | 6.4M (0.075)         | 0.170 (0.034)                              | 0.209 (0.037)                               |
| Mistral-7b [42]  | 6.8M (0.096)         | <b>0.250 (0.002)</b>                       | <b>0.304 (0.002)</b>                        |

Kendall correlation of 0.179 and a Spearman correlation of 0.220, better than Gemma-7b-it [31] that has similar amount of trainable parameters. As a result, we opted for Mistral as our backbone LLM.

## Chapter 7

# ReFINE: A Reward-Based Framework for Interpretable and Nuanced Evaluation of Radiology Report Generation

While the previous chapter tackled the challenge of evaluating overall report quality, this chapter presents a more interpretable evaluation framework with detailed sub-score analysis. Despite notable progress in automated radiology report generation (R2Gen), the evaluation of generated reports remains difficult due to the nuanced nature of medical texts. Conventional metrics frequently diverge from human judgment, overlooking specific shortcomings in report quality. To bridge this gap, we propose **ReFINE**, a novel evaluation framework that trains an Evaluation Model using a **margin-based reward enforcement loss**. Our method breaks down overall quality into **fine-grained sub-scores** defined by user-specified criteria, enhancing interpretability and diagnostic precision. By utilizing GPT-4 to synthesize diverse training pairs of accepted and rejected reports, we train a reward-guided model capable of delivering both detailed sub-scores and a comprehensive quality estimate, enabling **criterion-level evaluation**. Empirical results highlight ReFINE's stronger alignment with expert assessments, surpassing traditional metrics in selecting high-quality models. Its effectiveness is further confirmed on three expert-labeled datasets, including chest X-rays and multimodal reports across 9 imaging modalities, evaluated under two distinct scoring

standards.

## 7.1 Introduction

Automated radiology report generation (R2Gen), which involves creating free-text summaries of visual observations from radiographic images, has gained significant momentum in recent years [105, 57]. This sophisticated AI task requires deep understanding of complex clinical semantics, rendering both the generation and evaluation phases particularly difficult. Commonly used natural language generation (NLG) metrics such as BLEU [82] and METEOR [88] emphasize surface-level n-gram overlaps and often overlook lexical variation and structural diversity critical for medical interpretability. While BERTScore [124] improves on this by leveraging contextual embeddings for paraphrasing, clinical-specific metrics like CheXbert [92] and RadgraphF1 [41] target structured findings but still miss broader semantic nuances.

Despite these advancements, current metrics frequently fail to reflect human expert judgment [67]. Attempts like RadCliQ [121] combine several metrics using regression over human annotations to better mimic expert preferences. However, RadCliQ’s dependency on limited and costly human-labeled data hampers its extensibility. Meanwhile, the emergence of large language models (LLMs) such as GPT-4 [77] offers new opportunities for report evaluation. Yet, direct use of such models raises privacy concerns, entails substantial computational costs, and may not be sustainable for large-scale R2Gen applications.

To mitigate these issues, we introduce **ReFINE**, a new evaluation framework tailored for radiology report generation. Unlike previous metrics, ReFINE utilizes an LLM fine-tuned with a novel margin-based reward enforcement loss, enabling it to break down report quality into fine-grained sub-scores aligned with user-defined criteria. This enhances interpretability and allows users to pinpoint the root causes of low-quality outputs—such as errors in lesion location, severity misjudgments, or omitted findings. We leverage GPT-4’s strong evaluation ability to automatically create human-like training data, forming pairs of "accepted"

and "rejected" reports using our predefined rule. These samples are used to train our evaluation model to generate both criterion-specific sub-scores and a unified global score.

Our proposed **ReFINE** framework offers several advantages over current methods. Static metrics (e.g., NLG-based or clinical scores) show poor correlation with expert assessments and lack flexibility for customization. Among learned metrics, RadCliQ is limited by its linear combination structure and static feature space. Some LLM-based solutions such as G\_Rad [13], FineRadScore [37], RadFact [7], and CheX-prompt [12] depend on commercial online LLMs, which can introduce privacy risks. While MRScore [68] and GREEN [78] are the most relevant, they present different limitations. MRScore simplifies the scoring task by producing a single scalar reward, sacrificing transparency and interpretability. Without sub-score granularity, it becomes difficult to perform detailed error attribution or fine-grained analysis. In contrast, ReFINE explicitly outputs multiple sub-scores, offering greater transparency and more actionable diagnostic feedback. GREEN enhances interpretability via textual explanations but compromises scoring consistency due to its free-form structure. Moreover, it lacks a dedicated reward-based objective like ours, weakening its sensitivity to subtle differences in report quality. Compared with GREEN, our method yields higher **Kendall's Tau correlation with human evaluation** (0.75 vs. 0.64), achieves better **resource efficiency** (1x NVIDIA A6000 vs. 8x A100 GPUs), and provides faster **inference speed**.

Our main contributions are summarized as follows:

- (1) We present a novel strategy for fine-tuning LLMs to evaluate radiology reports at a fine-grained level. The proposed ReFINE model uses a custom loss function to produce human-aligned rewards and achieves superior correlation with expert evaluations, all while running efficiently offline.
- (2) ReFINE outputs both comprehensive quality scores and sub-scores, offering interpretability and actionable insights by revealing which specific criteria a report fails to meet.
- (3) We evaluate ReFINE on three expert-annotated datasets—ReXVal,

RaTE-Eval, and our in-house Rad-100—spanning multiple imaging modalities and evaluation rubrics. ReFINE consistently demonstrates stronger alignment with human ratings than previous evaluation metrics.

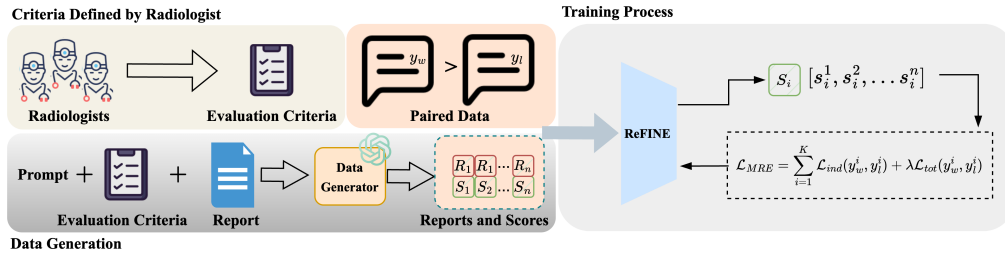


Figure 7.1: Training procedure of the ReFINE model.

## 7.2 Related work for LLM-based Evaluation Metrics

Recent efforts have explored the use of large language models (LLMs) for evaluating radiology reports. Several methods adopt prompt-based evaluation paradigms, utilizing commercial APIs such as GPT-4 or Claude-3 to assess various dimensions of report quality. For instance, **G-Rad Score** [13] prompts GPT-4 to identify six distinct types of clinically significant errors—including false prediction, omission, location and severity errors—and filters out clinically insignificant cases, offering targeted error analysis. **CheXPrompt** [12] did the same thing as G-RadScore.

**FineRadScore** [37] takes a more granular approach by prompting LLMs to perform line-level comparison between generated and reference reports. It explicitly classifies edits (insertion, deletion, substitution), assigns clinical severity scores, and categorizes errors into subtypes such as factual inaccuracy or clinical omission. This fine-grained annotation improves traceability but relies heavily on the language model’s internal consistency. **RadFact** [7] focuses on logical and spatial consistency between the generated and ground-truth report by computing logical precision and recall using LLaMA-70B-Instruct. It can optionally incorporate spatial grounding accuracy without relying on task-specific radiology models.

While these systems report promising results, particularly in terms of qualitative feedback, they suffer from several critical limitations. First, all rely on external APIs, raising data privacy and security concerns, especially in clinical applications. Second, they incur high inference costs and are subject to version drift and unpredictable API behavior. Third, their scoring mechanisms are often opaque, limiting reproducibility and trustworthiness. For example, although **GREEN** [78] provides natural language explanations to enhance transparency, it exhibits lower scoring accuracy (Kendall’s Tau of 0.64) and lacks a dedicated loss function for optimizing sub-score prediction.

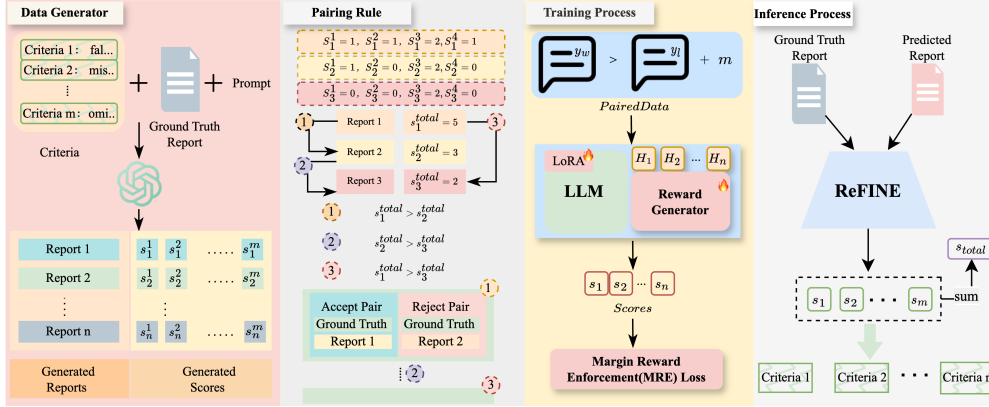
Moreover, when evaluated on the ReXVal benchmark, these LLM-based methods—despite their sophisticated prompting and error categorization—still fall short or only match the performance of **ReFINE** in aligning with expert human judgments.

In contrast, **ReFINE** is designed for fully offline deployment, eliminating the risk of data leakage associated with third-party APIs. While **MRScore** [68] also supports offline evaluation, it produces only a single reward score, which limits interpretability and hinders multi-faceted error analysis. **ReFINE** addresses these issues through a margin-based multi-subscore architecture, allowing for fine-grained, interpretable evaluation aligned with multiple clinical criteria. It is computationally efficient—trainable on a single A6000 GPU—and supports criterion-specific reward learning via a purpose-designed loss. This makes it a practical and trustworthy tool for real-world clinical report evaluation, bridging the gap between automatic scoring and radiologist-level review.

## 7.3 Method

In this section, we introduce **ReFINE**, a novel evaluation metric designed to closely reflect human judgment by offering both holistic scores and fine-grained sub-score breakdowns to enhance interpretability. To construct training data, we utilize GPT-4 to generate paired report samples based on a defined pairing strategy, as detailed in Algorithm 2. These pairs are then used to train a reward model optimized with a customized

loss objective. During inference, the model outputs criterion-specific sub-scores, which are aggregated to yield the final comprehensive score, as illustrated in Figure 7.2.



**Figure 7.2:** Overview of the framework for our model. The diagram is divided into four main parts: 1) **Data Generator:** Generates reports and corresponding scores based on specified criteria. 2) **Pairing Rule:** Demonstrates the scoring and pairing process using four criteria as an example. Reports are paired into "accepted" and "rejected" categories based on their total scores and margins. 3) **Training Process:** Utilizes paired data to train a reward model through a LoRA-based large language model (LLM) to optimize the MRE loss. 4) **Inference Process:** rate the predicted report by comparing it to the ground truth report, generating both sub-scores and total scores for evaluation.

### 7.3.1 Scoring Data Generation Pipeline

Recent research has highlighted GPT-4’s effectiveness in evaluating chest X-ray reports. When provided with well-defined evaluation criteria, **GPT-4 is capable of producing similarity assessments that align closely with human judgments**, as evidenced in both [18] and [68]. For instance, [18] reports that **GPT-4 attained a Kendall’s Tau of 0.735 when compared with radiologists’ annotations under the RadCliQ scoring scheme**. Similarly, in [68], **GPT-4 achieved a Kendall’s Tau correlation of 0.531 with expert ratings using the MRScore framework**. This variation in performance is likely attributable to the distinct emphases of the two scoring systems: RadCliQ mainly focuses on quantifying errors across individual sub-criteria, whereas MRScore additionally incorporates semantic-level

factors such as lesion characterization, overall report completeness, grammatical accuracy, and proper medical terminology. Motivated by these findings, we leverage GPT-4 to generate a large-scale scoring dataset, including both the reports and their associated scores, which is used to train our evaluation model. The complete procedure is described in detail below.

**Defining Scoring Criteria.** Numerous evaluation criteria have been explored in prior studies. To showcase the adaptability of our model under different assessment protocols, we adopt two distinct scoring systems in this work. The first is the RadCliQ scoring scheme introduced in [121], which categorizes errors into six types: (1) incorrect prediction of a finding, (2) omission of a relevant finding, (3) misidentification of a finding’s location or position, (4) erroneous severity assignment, (5) inclusion of a comparison not mentioned in the reference impression, and (6) exclusion of a comparison that indicates a temporal change from prior studies. This system primarily emphasizes clinical relevance by tallying the number of errors.

In contrast, the MRScore framework proposed in [68] incorporates both clinical and linguistic aspects. It defines seven core evaluation items grounded in domain knowledge and prior literature: “impression consistency,” “impression organs,” “lesion description,” “clinical history,” “completeness,” “grammar,” and “medical terminology.” Each item is assessed as a binary error (yes/no) and carries a distinct weight from the set {30, 20, 20, 10, 10, 5, 5}, which contributes to the item’s weighted penalty. The overall score is computed as:

$$\text{Total\_score} = 100 - \sum_{i=1}^7 S_i \times W_i,$$

where  $S_i$  denotes the binary error indicator for the  $i$ -th criterion and  $W_i$  is its assigned weight. Using these defined rules, GPT-4 can be prompted to evaluate reports in accordance with either scoring system, as detailed in the following subsection.

**Generating Scoring Training Dataset.** Given a specified scoring framework, we crafted prompts that explicitly reflect the evaluation criteria, enabling GPT-4 to perform report assessments in a manner consistent with human judgment (refer to the Appendix for a prompt example). Utilizing the GPT-4 API, we produced diverse-quality reports based on randomly sampled ground-truth entries from the MIMIC-CXR dataset.

For the RadCliQ system, approximately 8000 ground-truth reports were selected, each used to generate three corresponding GPT-4 reports with ascending error severity (containing 0–2, 3–4, and 5–6 errors, respectively). These reports were then labeled with both overall scores and individual error-type annotations. For the MRScore setting, we curated 1800 ground-truth reports, each leading to the generation of three GPT-4 outputs spanning different quality ranges (0–40, 40–70, and 70–100). Each of these samples was assigned both a total score and detailed scores across the seven MRScore criteria.

To ensure the reliability of the generated training data, we conducted a manual validation on a randomly chosen subset of 50 GPT-4-generated examples. Two board-certified radiologists (each with over 10 years of experience) independently reviewed and scored the reports. Based on consensus ratings, the accuracy across evaluation items was as follows: 0.90 for Impression, 0.98 for Impression Organ, 0.86 for Lesion Description, 0.92 for Clinical History, 0.98 for Completeness, and perfect agreement (1.0) for both Grammar and Medical Terminology.

It is important to emphasize that GPT-4-generated data was strictly used for training purposes, while the effectiveness of our evaluation model was assessed using several human-annotated benchmark datasets.

### 7.3.2 Reward Model

Building upon a pretrained language model backbone (e.g., LLaMA3 [96]), ReFINE is optimized using a tailored loss function applied to pairs of reports to ensure its reward predictions align closely with human judgments. Each pair originates from the same reference report but differs in quality, enabling the model to learn nuanced distinctions in report content. Through this training process, ReFINE is guided to assign higher scores to superior reports while simultaneously outputting criterion-specific sub-scores that provide interpretable feedback on distinct quality dimensions.

During inference, ReFINE produces individual rewards for each evaluation criterion, which are then aggregated to yield the final ReFINE score. To enhance the model’s responsiveness to subtle quality differences, we incorporate a dedicated margin for both per-criterion and overall scores, encouraging the model to better differentiate reports with fine-grained quality variations.

#### Model Input

Our model is trained using input pairs of reports along with their associated score margins. Each pair comprises an “accepted” and a “rejected” report, both derived from the same reference report, with the accepted version achieving a higher evaluation score. The pairing strategy is outlined in Algorithm 2, which selects the highest-scoring prediction for each ground truth by aggregating scores across multiple evaluation criteria. It then ranks the predicted reports and constructs pairs based on score differentials. These margins are calculated at both the overall and individual criterion levels, promoting both effective learning and interpretable report comparison.

#### Multi-Reward Generator

Our reward model is constructed upon the LLaMA3 [74] architecture, augmented with a multi-reward head to output the ReFINE score. LLaMA3

was chosen due to its strong language understanding capabilities, requiring only 6.8M trainable parameters out of its 7 billion total parameters. The multi-reward head consists of a linear projection layer that transforms the final hidden representation of LLaMA3 into an  $N \times 1$  vector, where  $N$  denotes the number of sub-scores defined by our evaluation criteria. To enable efficient adaptation, we apply Low-Rank Adaptation (LoRA) [36], facilitating fine-tuning with minimal parameter updates. The model is trained on paired “accepted” and “rejected” reports, allowing it to learn reward predictions aligned with quality differences captured by a scoring margin. Each sub-score corresponds to a specific criterion, and their combined sum yields the final ReFINE score that reflects the overall quality of the generated report.

---

**Algorithm 2** Report Pairing Rule with Margin Calculation

---

**Input:** Ground Truth Report  $x$ , Predicted Reports  $Y = \{y_1, y_2, \dots, y_n\}$ , Scoring System  $S$  with  $k$  criteria

**Output:** Accepted Reports  $Y_w$ , Rejected Reports  $Y_l$ , Margins (total and per criterion)

Initialize  $Y_w \leftarrow \emptyset, Y_l \leftarrow \emptyset, M \leftarrow \emptyset$

**for** each  $y_i$  in  $Y$  **do**

Compute  $S_{\text{tot}}(x, y_i) \leftarrow \sum_{j=1}^k S_j(x, y_i)$

**end for**

**for** each  $y_i$  in  $Y$  **do**

**for** each  $y_j$  in  $Y$  such that  $y_j \neq y_i$  and  $S_{\text{tot}}(y_j) < S_{\text{tot}}(y_i)$  **do**

Compute  $M_{\text{tot}} \leftarrow S_{\text{tot}}(y_i) - S_{\text{tot}}(y_j)$

Compute  $M_j \leftarrow S_j(y_i) - S_j(y_j)$  for all  $j = 1$  to  $k$

**if**  $M_{\text{tot}} > 0$  **then**

Add  $(x, y_i)$  to  $Y_w$

**else**

Add  $(x, y_i)$  to  $Y_l$

**end if**

Store  $M_{\text{tot}}$  and  $\{M_1, M_2, \dots, M_k\}$  in  $M$

**end for**

**end for**

**Return:**  $Y_w, Y_l, M = 0$

---

### Multi-Reward Learning.

The goal of our multi-reward framework is to emulate human judgment as reflected by GPT-4 rankings, by optimizing a learning function grounded in these preferences for radiology reports. The model is trained to distinguish and predict the more favorable report in each pair, effectively capturing nuanced quality differences. Rather than learning from a single holistic score, our formulation explicitly targets both the overall preference and the preferences across individual evaluation criteria. The loss function is detailed in Section 7.4. By incorporating both total and criterion-level score margins, the model is encouraged to distinguish variations in quality not only globally but also locally across sub-scores. The magnitude of each margin governs the corresponding penalty, enabling the reward model to adaptively assign scores that reflect both fine-grained and holistic quality differences.

## 7.4 Margin Reward Enforcement(MRE) Loss Function

Considering a pair of generated reports  $\langle y_w^i, y_l^i \rangle$ <sup>1</sup> corresponding to the same  $i$ -th ground truth report  $x^i$ , the accepted report  $y_w^i$  receives a higher ground truth score (denoted as  $s_w^i$ ), while the rejected report  $y_l^i$  receives a lower ground truth score ( $s_l^i$ ). Let  $s_w^{i,j}$  and  $s_l^{i,j}$  denote the  $j$ -th sub-score of  $s_w^i$  and  $s_l^i$ , respectively, where  $j = 1, \dots, N$ , and  $N$  is the number of sub-scores under a specific scoring system. Note that although the total score  $s_w^i$  is greater than  $s_l^i$ , an individual sub-score  $s_w^{i,j}$  may not necessarily be greater than  $s_l^{i,j}$ . Our goal is to train the model to distinguish ranking relationships at both the individual sub-score and total score levels, ensuring alignment with the ground truth ranking structure, formulated as follows.

<sup>1</sup>Here "w" stands for "win", indicating the accepted report, and "l" for "lose", indicating the rejected report.

### 7.4.1 Individual Reward Loss $\mathcal{L}_{\text{ind}}$

This loss focuses on correctly ranking the individual sub-scores for each scoring criterion.

$$\mathcal{L}_{\text{ind}}(y_w^i, y_l^i) = \frac{1}{N} \sum_{j=1}^N \left( \mathbb{1}(s_w^{i,j} \neq s_l^{i,j}) \text{ReLU}(-t_w(r_w^{i,j} - r_l^{i,j}) + t_w m^{i,j}) + (1 - \mathbb{1}(s_w^{i,j} \neq s_l^{i,j})) \text{ReLU}(|r_w^{i,j} - r_l^{i,j}| - c) \right), \quad (7.1)$$

- $t_w = 1$  if  $m^{i,j} > 0$ , otherwise  $t_w = -1$ .
- $m^{i,j} = s_w^{i,j} - s_l^{i,j}$  is the margin between sub-scores for the  $j$ -th criterion.
- $\mathbb{1}(s_w^{i,j} \neq s_l^{i,j})$  ensures this term applies only when the sub-scores differ.

Here  $r_w^{i,j}$  and  $r_l^{i,j}$  denote the  $j$ -th individual rewards assigned to the reports  $y_w^i$  and  $y_l^i$ , respectively. The margin between the total scores  $s_w^i$  and  $s_l^i$  is denoted by  $m^i = s_w^i - s_l^i$ , where  $m^i > 0$ . The individual "margin"  $m^{i,j} = s_w^{i,j} - s_l^{i,j}$  is not necessarily positive. The variable  $t_w$  acts as a flag:  $t_w = 1$  if  $m^{i,j} > 0$ , otherwise  $t_w = -1$ . The function  $\mathbb{1}(\cdot)$  is an indicator function, returning 1 when the event occurs and 0 otherwise.  $K$  is the total number of report pairs.

- If  $s_w^{i,j} > s_l^{i,j}$  (i.e.,  $m^{i,j} > 0$ ): A penalty is incurred if  $r_l^{i,j} > r_w^{i,j} - m^{i,j}$ , indicating that the "losing" report's reward exceeds the acceptable range relative to the "winning" report.
- If  $s_w^{i,j} < s_l^{i,j}$  (i.e.,  $m^{i,j} < 0$ ): A penalty is incurred if  $r_l^{i,j} < r_w^{i,j} - m^{i,j}$ , meaning the "losing" report's reward falls below the acceptable range.
- If  $s_w^{i,j} = s_l^{i,j}$  (i.e.,  $m^{i,j} = 0$ ): A penalty is incurred if  $|r_w^{i,j} - r_l^{i,j}| > c$ , ensuring that rewards remain sufficiently close when scores are equal.

### 7.4.2 Total Reward Loss $\mathcal{L}_{\text{tot}}$

This loss enforces a clear margin between the total scores of the winning and losing reports.

$$\mathcal{L}_{\text{tot}}(y_w^i, y_l^i) = \text{ReLU} \left( - \left( \sum_{j=1}^N r_w^{i,j} - \sum_{j=1}^N r_l^{i,j} \right) + m^i \right), \quad (7.2)$$

where  $m^i = s_w^i - s_l^i$  is the margin between total scores  $s_w^i$  and  $s_l^i$ .

- If  $\sum_j r_l^{i,j} \leq \sum_j r_w^{i,j} - m^i$ : No penalty, as the rejected report is within the acceptable range.
- If  $\sum_j r_l^{i,j} > \sum_j r_w^{i,j} - m^i$ : Penalty applied, as the rejected report exceeds the acceptable margin.

### 7.4.3 Margin Reward Enforcement (MRE) Loss

The overall loss,  $\mathcal{L}_{\text{MRE}}$ , combines these two terms: the individual reward loss  $\mathcal{L}_{\text{ind}}$  and the total reward loss  $\mathcal{L}_{\text{tot}}$ , balanced by the hyperparameter  $\lambda$ , an ablation of hyperparameter  $\lambda$  shows in Table 7.9. This combined loss ensures that the model learns to rank individual sub-scores correctly through  $\mathcal{L}_{\text{ind}}$ , while maintaining a sufficient margin between total scores via  $\mathcal{L}_{\text{tot}}$ . By minimizing  $\mathcal{L}_{\text{overall}}$ , the model effectively balances individual and total reward losses, enabling it to provide nuanced insights into the assessment results by optimizing both individual sub-scores and the total score.

$$\mathcal{L}_{\text{MRE}} = \sum_{i=1}^K \left( \mathcal{L}_{\text{ind}}(y_w^i, y_l^i) + \lambda \mathcal{L}_{\text{tot}}(y_w^i, y_l^i) \right), \quad (7.3)$$

The loss function consists of multiple components, including ReLU and absolute functions, whose gradients can be derived using their subgradient properties:

$$\frac{\partial \mathcal{L}_{\text{ind}}}{\partial r_w^{i,j}} = \begin{cases} -t_w, & \text{if } s_w^{i,j} \neq s_l^{i,j} \text{ and } r_w^{i,j} - r_l^{i,j} < 0 \\ 1, & \text{if } s_w^{i,j} = s_l^{i,j} \text{ and } |r_w^{i,j} - r_l^{i,j}| > c \\ 0, & \text{otherwise} \end{cases} \quad (4a)$$

$$\frac{\partial \mathcal{L}_{\text{tot}}}{\partial r_w^{i,j}} = \begin{cases} -1, & \text{if } \sum_{j=1}^N r_w^{i,j} - \sum_{j=1}^N r_l^{i,j} < m^i \\ 0, & \text{otherwise} \end{cases} \quad (4b)$$

## 7.5 Experiments and Result

All our models are trained by one A6000 GPU with 48 GB memory.

### 7.5.1 Datasets

We assess the performance of ReFINE on three datasets annotated by domain experts: ReXVal [122], RaTE-Eval [126], and Rad-100. While both ReXVal and RaTE-Eval utilize the RadCliQ scoring system, they differ significantly in scope—ReXVal focuses exclusively on chest X-rays (CXRs), whereas RaTE-Eval encompasses a broader spectrum, covering nine imaging modalities and 22 anatomical regions. In contrast, Rad-100 is limited to CXR data but employs the MRScore framework, offering an opportunity to examine ReFINE’s adaptability across distinct evaluation schemes.

**ReXVal** [122] is a publicly available benchmark dataset containing 200 candidate-reference pairs derived from 50 MIMIC-CXR studies (with four candidate reports per study). Each pair is annotated by six board-certified radiologists following RadCliQ’s six error categories, providing a standard for evaluating metric alignment with human expert ratings.

**RaTE-Eval** [126] includes 2,215 sentence-level report pairs from MIMIC-IV, annotated using RadCliQ by two radiologists with over five years of clinical experience. It covers diverse imaging types such as CT, MRI, and Ultrasound, and is organized into an 8:2 train/test split to support both model development and evaluation.

**Rad-100** is a dataset constructed under the MRScore evaluation framework (see Section 3.1), consisting of 100 generated reports from the R2Gen model, each aligned with a randomly chosen ground-truth report from MIMIC-CXR. Scoring was independently performed by three senior radiologists (each with over a decade of clinical experience), with majority voting used to finalize scores on a per-criterion basis. The total score was obtained by aggregating these sub-scores. Importantly, Rad-100 was not

used in training ReFINE <sup>2</sup>, making it a reliable benchmark for evaluating generalization to a different scoring rubric beyond RadCliQ.

## 7.5.2 Performance on ReXVal Dataset

**Correlation Analysis of Sub-criteria.** Table 7.1 shows ReFINE’s performance on the ReXVal dataset using the RadCliQ Scoring System. ReFINE demonstrates strong alignment with expert judgments across error categories, with overall Kendall’s Tau of 0.751 and Spearman correlation of 0.910. High scores in categories such as "False prediction of a finding" (Kendall: 0.680, Spearman: 0.842) and "Omission of a finding" (0.507, 0.673) indicate effective recognition of common radiological errors. Lower scores for "Incorrect location or position" (0.246, 0.327) suggest difficulty capturing subtle spatial details. Unlike single-score metrics [121, 123, 41], ReFINE’s sub-criteria output helps pinpoint specific model weaknesses, guiding targeted improvements.

**Table 7.1:** Human Correlations of ReFINE Subscores on ReXVal Dataset

| Criteria  | Kendall’s Tau <sup>↑</sup> | Spearman <sup>↑</sup> |
|---|----------------------------|-----------------------|
| - Score 1: False prediction of a finding                                      | 0.680                      | 0.842                 |
| - Score 2: Omission of a finding  | 0.507                      | 0.673                 |
| - Score 3: Incorrect location or position of a finding                        | 0.246                      | 0.327                 |
| - Score 4: Incorrect severity of a finding                                    | 0.443                      | 0.569                 |
| - Score 5: Mention of a comparison absent in the reference impression         | 0.433                      | 0.545                 |
| - Score 6: Omission of a comparison that notes a change from a previous study | 0.267                      | 0.345                 |
| Total   | 0.751                      | 0.910                 |

**Comparison with other metrics.** Table 7.2 presents the performance comparison of various evaluation metrics on the ReXVal dataset, measured using Kendall’s Tau and Spearman correlation based on total scores. In contrast to ReFINE, *most existing metrics lack the flexibility to support user-defined sub-criteria*, making sub-score level comparisons infeasible.<sup>3</sup>

<sup>2</sup>Rad-100 was held out entirely from the ReFINE training pipeline.

<sup>3</sup>GREEN offers error counts for each subcategory but does not report sub-score level correlations with human annotations. Furthermore, its overall agreement with human assessments is considerably lower than that of ReFINE.

The comparison includes traditional NLG metrics (BLEU-4 [82], ROUGE-L [61], METEOR [6], CIDEr [100]), clinical metrics (BERTScore [123], Rad-GraphF1 [41], and Semb\_score [121]), hybrid methods (RadCliQ [121]), and both offline (GREEN [78], RaTEScore [126]) and online LLM-based approaches (RadFact [7], CheXPrompt [12], G-Rad [13], FineRadScore [37]). ReFINE achieves the best overall correlation scores with human ratings (Kendall: 0.751, Spearman: 0.910). Although CheXPrompt and G-Rad report comparable Kendall scores (0.750), their reliance on proprietary GPT-based APIs introduces privacy concerns and hampers reproducibility.

Offline alternatives such as GREEN (0.640, 0.816) and RadCliQ-v1 (0.631, 0.816) underperform, with GREEN additionally requiring substantially higher computational resources—8 A100 GPUs for 12 training epochs, compared to ReFINE’s efficient setup of 1 A6000 GPU over just 4 epochs. Crucially, ReFINE uniquely supports customizable evaluation based on sub-criteria, a capability not provided by prior metrics. These findings underscore ReFINE’s superior alignment with expert ratings, enhanced interpretability, and computational efficiency.

### **7.5.3 Performance on RaTE-Eval Dataset**

We conduct training and evaluation on the RaTE-Eval dataset, which spans nine distinct imaging modalities. For comparison, we follow the protocol of [126], which reports only the Pearson correlation without providing sub-score level details. As presented in Table 7.3, our method achieves a Pearson correlation of 0.61, outperforming all previously reported evaluation metrics.

### **7.5.4 Performance on Rad-100 Dataset**

Given that the Rad-100 dataset adopts a binary scoring system—where each sub-criterion is labeled as 1 if an error exists and 0 otherwise (refer to Appendix for detailed definitions)—the final score is computed by weighting these binary indicators with predefined coefficients. Based on this structure, we assess the sub-criterion classification accuracy of

**Table 7.2:** Human Correlation Comparison of Evaluation Metrics on ReX-Val Dataset. *Gray part indicates online methods which directly use online model API.*

| Metric               | Kendall’s Tau $\uparrow$ | Spearman $\uparrow$ |
|----------------------|--------------------------|---------------------|
| BLEU-4 [82]          | 0.345                    | 0.475               |
| ROUGE-L [61]         | 0.491                    | 0.663               |
| METEOR [6]           | 0.464                    | 0.627               |
| CIDEr [100]          | 0.499                    | 0.664               |
| BertScore [123]      | 0.507                    | 0.677               |
| RadGraphF1 [41]      | 0.516                    | 0.702               |
| Semb_score [121]     | 0.494                    | 0.665               |
| RadCliQ-v1 [121]     | 0.631                    | 0.816               |
| GREEN [78]           | 0.640                    | -                   |
| RaTEScore [126]      | 0.527                    | -                   |
| <b>ReFINE (Ours)</b> | <b>0.751</b>             | <b>0.910</b>        |

*Results below are not strictly comparable because they are **online** model(e.g., GPT4).*

|                            |       |   |
|----------------------------|-------|---|
| RadFact (online) [7]       | 0.590 | - |
| CheXprompt (online) [12]   | 0.750 | - |
| G-Rad (online) [13]        | 0.750 | - |
| FineRadScore (online) [37] | 0.737 | - |

**Table 7.3:** Human Correlation on the Multimodal RaTE-Eval Dataset (Sentence-level correlation).  $\dagger$  indicates values directly cited from the RaTEScore paper; others are reproduced.

| Metric                   | BLEU           | ROUGE          | METEOR         | CIDEr          | BERTScore      | RadGraph       | RadCliQ        | RaTEScore      | ReFINE (Ours) |
|--------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------------|
| Pearson $\uparrow$       | 0.27 $\dagger$ | 0.34 $\dagger$ | 0.39 $\dagger$ | 0.25 $\dagger$ | 0.40 $\dagger$ | 0.44 $\dagger$ | 0.46 $\dagger$ | 0.54 $\dagger$ | <b>0.61</b>   |
| Spearman $\uparrow$      | 0.23           | 0.21           | 0.33           | 0.28           | 0.35           | 0.42           | 0.41           | -              | <b>0.59</b>   |
| Kendall’s Tau $\uparrow$ | 0.16           | 0.15           | 0.24           | 0.20           | 0.25           | 0.31           | 0.30           | -              | <b>0.45</b>   |

**Table 7.4:** Human Correlation (with p-values) on RaTE-Eval (Sentence-level, 9 modalities)

| Metric              | Score 1         | Score 2        | Score 3         | Score 4         | Score 5         | Score 6        |
|---------------------|-----------------|----------------|-----------------|-----------------|-----------------|----------------|
| Spearman $\uparrow$ | 0.41 (3.85e-19) | 0.15 (3.94e-2) | 0.45 (9.66e-25) | 0.32 (1.18e-11) | 0.25 (1.71e-10) | 0.11 (2.61e-2) |
| Kendall $\uparrow$  | 0.30 (5.99e-18) | 0.12 (3.79e-2) | 0.34 (7.03e-23) | 0.27 (1.54e-11) | 0.20 (3.08e-10) | 0.09 (2.66e-2) |
| Pearson $\uparrow$  | 0.41 (1.51e-18) | 0.15 (1.79e-2) | 0.44 (2.20e-22) | 0.37 (1.42e-14) | 0.23 (1.33e-9)  | 0.10 (7.33e-2) |

our model, as shown in Table 7.5. The overall score comparisons across evaluation metrics are summarized in Table 7.6. ReFINE continues to demonstrate strong performance, achieving a Kendall’s Tau of 0.230 and a Spearman correlation of 0.293, both of which surpass the results of competing metrics. The associated p-values ( $p = 0.003$  for both) affirm the

statistical significance of ReFINE’s alignment with expert annotations.

**Table 7.5:** Subscore Accuracies on Rad-100 dataset. ‘Imp. Cons.’ stands for Impression Consistency, ‘Imp. Org.’ for Impression Organ, ‘Desc. Les.’ for Description of Lesion, ‘Clin. Hist.’ for Clinical History, ‘Comp.’ for Completeness, ‘Gram.’ for Grammar, and ‘Med. Term.’ for Medical Terminology.

| Sub-criteria | Imp. Cons. | Imp. Org. | Desc. Les. | Clin. Hist. | Comp. | Gram. | Med. Term. |
|--------------|------------|-----------|------------|-------------|-------|-------|------------|
| Accuracy     | 0.589      | 0.730     | 0.770      | 0.410       | 0.380 | 0.980 | 0.720      |

**Table 7.6:** Human Correlation on Rad-100 Dataset. @B = BLEU-4, @R = ROUGE-L, @M = METEOR, @C = CIDEr, @BS = BERTScore, @RG = RadGraphF1, @Semb = semb\_score, @RC = RadCliQ-v1

| Metric         | @B   | @R   | @M   | @C   | @BS  | @RG  | @Semb | @RC  | Ours        |
|----------------|------|------|------|------|------|------|-------|------|-------------|
| Kendall’s Tau↑ | 0.07 | 0.16 | 0.11 | 0.04 | 0.13 | 0.09 | 0.01  | 0.08 | <b>0.23</b> |
| Spearman↑      | 0.05 | 0.12 | 0.08 | 0.03 | 0.09 | 0.06 | 0.01  | 0.06 | <b>0.29</b> |

### 7.5.5 Ablation study

**Loss Components.** Our proposed objective consists of two components: the individual reward loss  $L_{ind}$  and the aggregate reward loss  $L_{tot}$ . An ablation study analyzing the impact of each component is provided in Table 7.8. Results indicate that using only  $L_{tot}$  for sub-score prediction reduces the Kendall’s Tau from 0.751 to 0.740 for the final score (the sum of sub-scores). Conversely, training solely with  $L_{ind}$  results in a drop to 0.738. These findings underscore the importance of combining both terms, with  $L_{tot}$  serving as a valuable regularization term.

**Hyperparameter Settings.** Our loss formulation includes two hyperparameters. The parameter  $c$  is a small positive constant (set to  $1e-2$ ) used for thresholding when comparing  $r_w$  and  $r_l$ . The weighting factor  $\lambda$  modulates the contribution of  $L_{ind}$  and  $L_{tot}$ . Table 7.9 shows an ablation on  $\lambda$ , revealing that model performance remains robust across a range of reasonable values. Notably, even under varying  $\lambda$ , our method consistently surpasses traditional evaluation metrics in terms of alignment with human ratings.

**Performance Comparison of Different LLM backones** In Table 7.7, we present a comparative analysis of different LLM backbones. Notably, LLaMA3 demonstrates superior performance with a medium size of trainable parameters. To ensure the scoring system is easily deployable, we focused on models with 7 billion parameters in total or fewer.

**Table 7.7:** Ablation Study of LLM Backbones on ReXVal Dataset

| Model          | Trainable Params (%) | Kendall's Tau ( $\uparrow$ ) | Spearman ( $\uparrow$ ) |
|----------------|----------------------|------------------------------|-------------------------|
| LLaMA3 [74]    | 6.8M (0.090)         | 0.751                        | 0.910                   |
| Vicuna-7b [19] | 8.4M (0.127)         | 0.738                        | 0.901                   |
| Meditron [14]  | 8.4M (0.127)         | 0.709                        | 0.880                   |
| Gemma-7b [31]  | 6.4M (0.075)         | 0.707                        | 0.876                   |
| Qwen1.5-7b[4]  | 8.4M (0.110)         | 0.684                        | 0.858                   |
| Phi-2 [59]     | 5.3M (0.196)         | 0.591                        | 0.784                   |

**Table 7.8:** Ablation Study on Different Loss Terms

| $\mathcal{L}_{\text{tot}}$ | $\mathcal{L}_{\text{ind}}$ | Spearman ( $\uparrow$ ) | Kendall's Tau ( $\uparrow$ ) |
|----------------------------|----------------------------|-------------------------|------------------------------|
|                            |                            | 0.319                   | 0.215                        |
| ✓                          |                            | 0.899                   | 0.740                        |
|                            | ✓                          | 0.899                   | 0.738                        |
| ✓                          | ✓                          | <b>0.910</b>            | <b>0.751</b>                 |

**Table 7.9:** Varying  $\lambda$  Values Ablation

| $\lambda$                    | 0.5   | 0.8   | 1.0          | 1.2   | 2.0   | 3.0   |
|------------------------------|-------|-------|--------------|-------|-------|-------|
| Kendall's Tau ( $\uparrow$ ) | 0.743 | 0.746 | <b>0.751</b> | 0.740 | 0.735 | 0.729 |

## 7.5.6 Qualitative Analysis

A qualitative illustration is presented in Figure 7.3, highlighting the consistency between ReFINE scores and human assessments under the Rad-CliQ scoring framework. In this example, the generated report misrepresents the severity of the "left pleural effusion" (marked in red), which results in a high sub-score from ReFINE under the category "incorrect

severity of a finding"—in agreement with the human annotation. Furthermore, the report incorrectly refers to a "right pleural effusion", triggering a high score for "incorrect location/position of a finding", a discrepancy also identified by human evaluators. Lastly, the omission of the "left retrocardiac opacification" leads both ReFINE and human raters to assign a score of 1.0 for the "false prediction of a finding" category.

|   |              |              |   |              |              |
|---|--------------|--------------|---|--------------|--------------|
| Ground truth: <b>Left retrocardiac opacification</b> could be atelectasis or infection. Pulmonary vascular congestion without evidence of Interstitial edema. <b>Possible small left pleural effusion</b> |              |              | Predict: <b>Moderate left pleural effusion</b> with underlying atelectasis noting infection would also be possible. Pulmonary vascular congestion and <b>probable small right</b> pleural effusion as well. |              |              |
| Criteria  | ReFINE       | Human        | Criteria  | ReFINE       | Human        |
| 1) False prediction in predicted report   | <b>1.000</b> | <b>1.000</b> | 4) Incorrect severity of a finding  | <b>0.784</b> | <b>0.833</b> |
| 2) Omission of a finding  | 0.012        | 0.000        | 5) Mention of a comparison not present in the refere  | 0.000        | 0.000        |
| 3) Incorrect location/position of a finding   | <b>0.263</b> | <b>0.167</b> | 6) Omission of a comparion describing a change from a previous study  | 0.227        | 0.000        |

**Figure 7.3:** An visual example of ReFINE from ReXVal Dataset. The highlighted sentences in reports and their corresponding scores share the same colors.

## Chapter 8

# Conclusion and Future Work

### 8.1 Conclusion

In this thesis, we have systematically explored the design, training, and evaluation of medical vision-language models (VLMs), with the goal of advancing automated understanding and generation of clinically relevant information from medical images. We investigated three fundamental tasks in this domain: Visual Question Answering (VQA), Radiology Report Generation (RRG), and Visual Grounding (VG). Furthermore, we emphasized the critical importance of evaluation, proposing novel approaches that better align with human judgment and clinical needs.

To address the limitations of existing VQA systems, we proposed **Q2ATransformer**, a semi-open framework that unifies the benefits of classification- and generation-based models. By introducing learnable answer embeddings and a transformer-based decoding strategy, Q2A-Transformer enables richer semantic reasoning while preserving answer stability, significantly improving performance on both close-ended and open-ended medical questions.

For radiology report generation, we introduced **SAT-RRG**, a self-adaptive training framework that leverages large language models (LLMs) to detect semantically incorrect tokens during training. By incorporating two custom-designed loss functions—CTAL and ETAPL—our method enables token-level feedback without requiring human annotations, effectively improving both factual correctness and report coherence.

We also conducted a comprehensive evaluation of **GPT-4V**, a state-of-the-art vision-language model, across the tasks of radiology report generation, VQA, and visual grounding. Through a combination of quantitative analysis, human evaluation, and case studies, we found that GPT-4V exhibits promising general capabilities in medical domains, particularly in RRG and VQA, but still struggles with spatial localization in VG. These findings underscore both the potential and the current limitations of general-purpose VLMs in clinical contexts.

Recognizing the inadequacy of existing evaluation metrics in capturing semantic and clinical correctness, we proposed two new scoring frameworks. First, **MRScore** is an LLM-based metric trained with GPT-generated pseudo-labeled data that demonstrates higher alignment with expert assessments while requiring no manual annotations. Second, **ReFINE** extends this idea by introducing fine-grained, criterion-specific scores trained via a novel margin-based reward enforcement loss. ReFINE enables interpretable, clinically meaningful evaluations, outperforming prior metrics in both accuracy and flexibility.

Together, the contributions of this thesis provide a unified perspective on how large-scale vision-language models can be developed and evaluated for medical applications. We not only propose new model architectures and training strategies but also advance the state-of-the-art in evaluation methodology, which is essential for trustworthy deployment in clinical workflows.

## 8.2 Future Work

While our work makes significant strides in improving the quality and evaluation of medical vision-language models, several promising directions remain for future exploration:

- **Towards end-to-end multimodal reasoning agents.** Future research can explore the development of *holistic end-to-end medical AI agents* that integrate multiple vision-language tasks, including Visual Question Answering (VQA), Radiology Report Generation (RRG), and

Visual Grounding (VG), into a single interactive framework. Such agents would function as intelligent clinical assistants, capable of supporting multi-turn, context-aware dialogues that evolve over time. For example, a physician might ask about suspicious lesions, request summarization of findings, and then follow up with temporal comparisons, all within the same session. Achieving this will require unified input-output modeling, task-switching capabilities, and dialogue memory integration, ensuring that the agent maintains clinical consistency and reasoning across turns.

- **Incorporating temporal and multi-modal data.** A significant portion of clinical decision-making relies on temporal trends and cross-modal correlations, such as comparing follow-up scans or linking imaging with lab results, notes, or genetic data. To better mimic real clinical workflows, models should be extended to handle temporally, ordered image sequences (e.g., progression of nodules across CT scans) and heterogeneous data types (e.g., combining imaging with EHR or pathology reports). This would require advancements in multi-stream fusion architectures, temporal attention mechanisms, and data harmonization techniques, as well as novel benchmarks that emphasize disease evolution and treatment response tracking.
- **Better grounding and explainability.** While models like ReFINE provide sub-score-level interpretability, a key limitation remains in generating clinician-understandable justifications—particularly visual or textual rationales tied to specific outputs. Future work should aim to strengthen the grounding between image regions and generated text, possibly via attention visualization, counterfactual explanations, or gradient-based saliency maps. Moreover, enabling models to output explanatory narratives alongside predictions (e.g., “This sentence is based on the opacity observed in the right lung field”) could significantly enhance clinical trust, auditability, and diagnostic value, especially in high-stakes settings.
- **Evaluation beyond alignment: calibration and safety.** Most current evaluation methods focus on alignment with human ratings

(e.g., via correlation metrics), but fail to assess model reliability under uncertainty. Future metrics should quantify how well a model calibrates its confidence, especially for ambiguous or rare cases, and assess error severity based on clinical impact. For example, hallucinating a pneumonia diagnosis is far riskier than omitting a benign rib fracture. Research should focus on building safety-aware scoring frameworks, incorporating uncertainty estimation, risk stratification, and outlier detection, enabling both automated triage and fallback to human review when necessary.

- **Lightweight and domain-adaptable models.** The deployment of large-scale medical models in real-world hospitals is often hindered by compute, latency, and privacy constraints. Future efforts should investigate parameter-efficient adaptation methods (e.g., adapters), knowledge distillation, and quantization to compress models while preserving accuracy. Moreover, models must be robustly adaptable to domain shifts, such as transitioning from tertiary-care hospitals to rural clinics or between different imaging vendors. Developing modular, plug-and-play architectures that can be fine-tuned with minimal labeled data will facilitate broader, equitable access to medical AI technologies.
- **Human-in-the-loop supervision.** Despite recent advances in self-supervised and instruction-tuned learning, radiologist-in-the-loop training remains underutilized. Future systems should actively incorporate clinician preferences and corrections via techniques like preference modeling, online reinforcement learning with human feedback (RLHF), or uncertainty-guided active learning. This interactive supervision loop can improve model alignment with clinical goals, accelerate error correction, and foster user trust. Additionally, designing intuitive interfaces for radiologists to provide feedback (e.g., through clickable heatmaps or report editing GUIs) will be essential for seamless integration into clinical workflows.

In conclusion, we hope that the methods and insights presented

in this thesis will serve as a foundation for future research on safe, effective, and interpretable medical vision-language AI systems, and ultimately contribute to better tools for supporting clinical care and decision-making.



## Appendix A

# Chapter 4 Appendix

### A.1 Prompt for Error Token Annotation

**Instruction.** Your task is to compare a *predicted report* with a *ground truth* medical report and identify specific tokens or phrases in the predicted report that are semantically incorrect or conflict with the ground truth. For this task:

1. **Incorrect tokens or phrases** are defined as parts of the predicted report that have a different meaning or contradict the ground truth.
2. Use a **matching pair of <e> and </e>** to wrap only the incorrect tokens or phrases in the predicted report.
3. Ensure that every error is marked precisely. Do not mark entire sentences—only the specific parts that are incorrect.
4. Preserve the structure of the predicted report. Do not split paragraphs or reformat the text.

#### Examples.

- **Example 1:**
  - Ground truth: *The lungs are clear and hyperinflated.*
  - Predicted report: *The lungs are clear and hyperinflation is present.*
  - Analysis: The predicted report matches the ground truth report.

- Output: The lungs are clear and hyperinflation is present.
- **Example 2:**
  - Ground truth: *Findings: The lungs are low in volume. No focal airspace consolidation to suggest pneumonia.*
  - Predicted report: *Findings: The lungs are within normal volume. Focal consolidation is noted in the right lower lobe, concerning for pneumonia.*
  - Analysis: The predicted report incorrectly states *within normal volume* and adds *Focal consolidation*, which conflicts with the ground truth.
  - Output: Findings: The lungs are `<e>within normal volume` `</e>`. Focal consolidation `<e>is noted` `</e>` in the right lower lobe, concerning for pneumonia.
- **Example 3:**
  - Ground truth: *Impression: There is evidence of acute cardiopulmonary process.*
  - Predicted report: *Impression: No acute cardiopulmonary process.*
  - Analysis: The predicted report incorrectly negates the cardiopulmonary process described in the ground truth.
  - Output: Impression: `<e>No acute` `</e>` cardiopulmonary process.
- **Example 4:**
  - Ground truth: *The patient has a 12-cm calcified granuloma unchanged from the prior study.*
  - Predicted report: *The patient has a mass in the lower lung field.*
  - Analysis: The predicted report incorrectly describes a *mass in the lower lung field*, which conflicts with the *12-cm calcified granuloma* in the ground truth.

- Output: The patient has a `<e>mass` in the lower lung field`</e>`.

Analyze the following reports and return the predicted report with incorrect tokens or phrases wrapped in matching pairs of `<e>` and `</e>`. Focus only on semantic differences, and ensure no extra modifications are made to the predicted report.



## Appendix B

### Chapter 5 Appendix

#### B.1 Details of Prompt Settings

To elucidate our prompt design and selection process, we present illustrative cases for better comprehension. We showcase instances involving both zero-shot and few-shot prompts to elucidate the distinction. In the zero-shot scenario, the model relies solely on its inherent training and knowledge base. Conversely, the few-shot scenario incorporates specific examples, enhancing the model's understanding and response accuracy. Our few-shot prompt encompasses diverse cases utilizing both normal and abnormal examples, showcasing radiology images with normal and abnormal findings. In all prompts, we prompt GPT-4V to assume the role of a professional radiologist. Additionally, we explicitly instruct it to generate both the impression and findings sections.

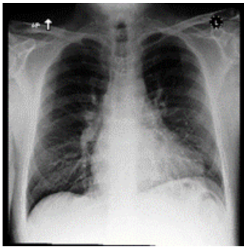
##### B.1.1 Zero-shot prompt

Figure B.1 showcases a zero-shot prompt example. We did not add any additional information to the text prompt.

##### B.1.2 Few-shot prompt

We add two example reports to the prompt while exploring three different combinations: (1) exclusively using normal examples, (2) exclusively using abnormal examples, (3) combining one normal and one abnormal example. The example reports are given in Figure B.2.

Zero-shot Prompt

|  |  |
|--|--|
| <p><b>Text-Prompt:</b></p> <p>You are a professional radiologist. I will provide you with some chest X-ray images. Please generate a radiology report based on the provided image. The report needs to contain both the impression and findings sections. You should only generate reports and no other unnecessary information.</p>   | <p><b>Image:</b></p>  |
| <p><b>GPT-4V:</b></p> <p>Impression: Normal chest radiograph with no acute cardiopulmonary abnormalities.<br/> Findings:<br/> <b>Lungs:</b> Clear bilaterally. No focal opacity, consolidation, or mass identified.<br/> <b>Cardiomediastinal silhouette:</b> Normal size and contour. No mediastinal widening.<br/> <b>Bones:</b> Intact without any appreciable fractures or bony lesions.<br/> <b>Diaphragm:</b> Intact with no evidence of diaphragmatic hernia.<br/> <b>Soft tissues:</b> Unremarkable.</p> |  |
| <p><b>Ground Truth:</b></p> <p>impression: No acute intrathoracic process.<br/> Findings:<br/> The lungs are well expanded and clear. The cardiomeastinal silhouette, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. Sternal wires are intact.</p>   |  |

**Figure B.1:** Zero-shot prompt. No additional information was provided to GPT-4V.

**Few-shot normal examples prompt** In this prompt method, we curated reports from two normal samples within the MIMIC-CXR training set. To ensure comprehensiveness, we specifically chose reports with rich content.

**Few-shot abnormal examples prompt** In this prompt method, we carefully chose two reports originating from abnormal samples within the MIMIC-CXR training set.

**Few-shot mixed-example prompt** In this prompt method, we chose one normal and one abnormal report from the MIMIC-CXR training set. The sequence in which these two examples are presented is not anticipated to significantly impact the generated results. In this specific experiment, we positioned the abnormal report before the normal one.

| <b>Few-shot abnormal examples prompt</b>  |
|---|
| <p><b>Case 1:</b><br/>           Impression: <b>Mediastinal widening</b> more than expected given procedure, concerning for substantial <b>mediastinal hemorrhage</b>. <b>Left chest tube</b> in place with no <b>pneumothorax</b>.<br/>           Findings: <b>Mediastinum</b> is widened and is more than expected status post bisegmentectomy concerning for substantial <b>mediastinal hemorrhage</b> with indentation of the <b>left tracheal</b> wall contour. Patient is status post <b>left upper lobe bisegmentectomy</b> with <b>left chest tube</b> in place. Bilateral low lung volumes. No <b>pneumothorax</b>. <b>Pleural effusions</b> little if any. <b>Cardiac</b> size appears enlarged and likely is exaggerated by low lung volumes.</p> <p><b>Case 2:</b><br/>           Impression: <b>Focal consolidation</b> at the left lung base, possibly representing <b>aspiration</b> or <b>pneumonia</b>. <b>Central vascular</b> engorgement.<br/>           Findings: PA and lateral views of the chest provided. The lungs are adequately aerated. There is a <b>focal consolidation</b> at the left lung base adjacent to the <b>lateral hemidiaphragm</b>. There is <b>mild vascular engorgement</b>. There is bilateral apical <b>pleural</b> thickening. The <b>cardiomediastinal silhouette</b> is remarkable for aortic arch calcifications. The heart is top normal in size.</p> |
| <b>Few-shot normal examples prompt</b>  |
| <p><b>Case 1:</b><br/>           Impression: No acute <b>cardiopulmonary</b> process.<br/>           Findings: There is no <b>focal consolidation</b>. <b>Pleural effusion</b> or <b>pneumothorax</b>, Bilateral nodular <b>opacities</b> that most likely represent <b>nipple</b> shadows. The <b>cardiomediastinal silhouette</b> is normal, Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable.</p> <p><b>Case 2:</b><br/>           Impression: No evidence of <b>pneumonia</b>.<br/>           Findings: The lung volumes are normal. Normal <b>hilar</b> and <b>mediastinal</b> structures. No <b>pneumonia</b>, no <b>pulmonary edema</b>. No <b>pleural effusions</b>. Status post CABG with aligned median sternotomy wires and normal location of surgical clips. Status post right lung surgery with surgical material seen.</p>  |
| <b>Few-shot mixed examples prompt</b>   |
| <p><b>Case 1:</b><br/>           Impression: <b>Retrocardiac opacity</b>, potentially <b>atelectasis</b>, though <b>infection</b> cannot be excluded in the correct clinical setting. Persistent elevation of the <b>right hemidiaphragm</b> with <b>right basilar atelectasis</b>.<br/>           Findings: Mild <b>cardiomegaly</b> with a <b>left ventricular predominance</b> is re-demonstrated. The <b>mediastinal and hilar contours</b> are unchanged. Pulmonary <b>vasculature</b> is normal. Elevation of the <b>right hemidiaphragm</b> is again noted with associated <b>right basilar atelectasis</b>. Retrocardiac <b>patchy opacity</b> may reflect <b>atelectasis</b> though <b>infection</b> is not excluded in the correct clinical setting. No <b>pleural effusion</b> or <b>pneumothorax</b> is detected. <b>S-shaped rotary scoliosis</b> of the <b>thoracolumbar spine</b> is again noted.</p> <p><b>Case 2:</b><br/>           Impression: No evidence of <b>pneumonia</b>.<br/>           Findings: Frontal and lateral views of the chest demonstrate no focal <b>consolidation</b>. Except for linear <b>atelectasis</b> in the right lower lung, the lungs are clear. The <b>cardiomediastinal</b> and <b>hilar contours</b> are normal. There is no <b>pneumothorax</b> or <b>pleural effusion</b>. Pleural surfaces are normal.</p>  |

**Figure B.2:** Example reports in prompts: Three pairs of different example reports in few-shot prompt settings. We added these example reports to few-shot prompts to help GPT-4V generate radiology reports.



## Appendix C

# Chapter 7 Appendix

### C.1 Labeled data prompt design and results

#### Prompt for RadCliQ Scoring System

Given a ground truth diagnostic report, generate three similar predicted reports. These predicted reports should be rated based on the following error-counting rules. The reports should have three levels of errors:

- 1st level has 0 or 1 errors,
- 2nd level has 2, 3 or 4 errors,
- 3rd level has 5 or 6 errors.

Error counting rule:

Given a ground truth diagnostic report and a predicted report, score the predicted report based on these error categories, each with significant and non-significant errors:

- False prediction of a finding
- Omission of a finding
- Incorrect location/position of a finding
- Incorrect severity of a finding
- Mention of a comparison not present in the reference impression
- Omission of a comparison describing a change from a previous study

For each error category, assign 1 point for significant errors and 1 point for non significant errors. The final score is the sum of these points.

Please generate three predicted reports for the given ground truth report. After that, score these three pairs (each predicted report with the ground truth) based on the error categories mentioned.

ground\_truth\_report: impression: no acute cardiopulmonary process.  
Cardiomegaly findings: frontal and lateral chest radiographs demonstrate marked cardiac enlargement unchanged compared to. Lungs are fairly well-aerated without focal consolidation pleural effusion or pneumothorax. The visualized upper abdomen is unremarkable.

Output format:

```
"json
{
  "ground_truth_report": "your_ground_truth_report_here",
  "predicted_reports": [
    {
      "predicted_report": "your_predicted_report_1",
      "errors": {
        "false_prediction": "your_score",
        "omission": "your_score",
        "incorrect_location": "your_score",
        "incorrect_severity": "your_score",
        "comparison_not_present": "your_score",
        "omission_of_comparison": "your_score"
      },
      "total_score": "your_total_score_1"
    },
    {
      "predicted_report": "your_predicted_report_2",
      "errors": {
        "false_prediction": "your_score",
        "omission": "your_score",
        "incorrect_location": "your_score",
        "incorrect_severity": "your_score",
        "comparison_not_present": "your_score",
        "omission_of_comparison": "your_score"
      },
      "total_score": "your_total_score_2"
    },
    {
      "predicted_report": "your_predicted_report_3",
      "errors": {
```

```

    "false_prediction": "your_score",
    "omission": "your_score",
    "incorrect_location": "your_score",
    "incorrect_severity": "your_score",
    "comparison_not_present": "your_score",
    "omission_of_comparison": "your_score"
  },
  "total_score": "your_total_score_3"
}
]
}

```

Please directly output the json file, no other contents

### Prompt for MRscore Scoring System

You are a skilled radiologist tasked with following task:

First Task:

By providing you with a "ground truth" report, generate three different reports, each with a score falling within specified score ranges. The scoring rules are detailed under the second task.

The score ranges for the reports are as follows: the first report scores between 0 to 40 points, the second report scores between 40 to 70 points, and the third report scores between 70 to 100 points.

Please generate a wider dispersion of scores

Second Task:

Evaluate radiology reports,

The three generated reports mentioned above are referred to as "predicted reports," and each is paired with the given "ground truth" report. Therefore, we will evaluate the three pairs of reports based on the following rules.

To achieve this objective, we compare the predicted report with the ground truth report to identify discrepancies between them. These discrepancies are defined according to the 'Error category' described in the table below, with each error assigned a specific weight. Upon the detection of an error, the weight is deducted from the total score of 100 according to the corresponding rule as follows. Analysis why

The scoring rule is:

- Check the predicted report for the presence of error items listed in the table below. Each item in the table needs to be checked, and if an error item is found, locate the corresponding score for this error item in the table and note it down, subtracting it from 100.
- Based on all the errors found, calculate all the error scores to get the total score, which means subtracting all existing error scores from 100.
- For the first error item under 'impression consistency', if there is no impression section in the ground truth, then this item does not count towards the score. Skip it with no score subtraction and proceed to analyze the other items in the table.
- Please generate the score and the analysis separately

Please format the result in a JSON format

| Error Category            | Description  | Score |
|---------------------------|--|-------|
| Impression consistency    | The impression shows normal or abnormal  | 30    |
| Impression Organ correct  | Is Lesion related Anatomical organ   | 20    |
| Description of Lesion     | Check the correctness of lesion location, lesion size, lesion opacity, Cardiovascular size, bone integrity | 20    |
| Clinical History          | Check the correctness of Operation history, treatment, family history                                      | 10    |
| Completeness truth report | Conclude all information in ground   | 10    |
| Grammar                   | Vocabulary spelling, fluently  | 5     |
| Medical Terminology       | Non-medical related terminology  | 5     |

Final Score = 100 - sum(Error Weight)

The given ground truth report is:

{content}

Output format example is as follows, if there is an error in the above rule, mark the corresponding score in the scoring part of the JSON

Predicted Report 0-40 indicates the quality of the generated report falls within the score bracket[0,40] For example 35

Predicted Report 40-70 indicates the quality of the generated report falls within the score bracket[40,70] For example 60

Predicted Report 70-100 indicates the quality of the generated report falls within the score bracket[70,100] For example 85

```

{{
  "Ground Truth Report": {content},
  "Predicted Report 0-40": {{
    "Ground Truth Report": "",
    "Predicted Report": "",
    "Scoring": {{}},
    "Analysis": {{}},
    "Final Score":
  }},

  "Predicted Report 40-70": {{
    "Ground Truth Report": "",
    "Predicted Report": "",
    "Scoring": {{}},
    "Analysis": {{}},
    "Final Score":
  }},

  "Predicted Report 70-100": {{
    "Ground Truth Report": "",
    "Predicted Report": "",
    "Scoring": {{}},
    "Analysis": {{}},
    "Final Score":
  }}
}}
```

## C.2 Labeled data

**Labeled data for RadCliQ Scoring System** In our study, we using LLM generated the labeled data as follows:

```

"ground_truth_report": "Impression: Bibasilar atelectasis. Tortuous
  aorta with likely ascending
aortic aneurysm, unchanged, which can better be assessed with contrast-
  enhanced CT. Emphysema.
Findings: Heart size is normal. Markedly tortuous thoracic aorta is
  again noted with dilatation
of the ascending aorta concerning for aneurysm, overall unchanged. The
  pulmonary vascularity
is not engorged. Calcified granulomas are noted bilaterally, the largest
  of which is in the
left upper lobe, unchanged. Streaky linear opacities are again noted
  within the lung bases
compatible with atelectasis. Lungs are hyperinflated with emphysematous
  changes again noted.
No focal consolidation, pleural effusion or pneumothorax is present. No
  acute osseous
abnormality is seen.",
"predicted_reports": [
{
  "predicted_report": "Impression: Mild bibasilar atelectasis. Aorta is
    tortuous with a
    possible ascending aortic aneurysm, showing progression, which should
    be evaluated with
    contrast-enhanced CT. Mild emphysema noted. Findings: Heart size
    appears slightly enlarged.
    The thoracic aorta shows tortuosity without clear evidence of aneurysm.
    No significant
    change in pulmonary vascularity. Multiple calcified granulomas are
    observed, with a notable
    one in the right lower lobe. Linear opacities suggest atelectasis at
    the lung bases. Lungs
    show signs of hyperinflation, indicative of emphysema. There is no
    evidence of consolidation,
    pleural effusion, or pneumothorax. No bone abnormalities detected.",
  "errors": {
    "false_prediction": "2",
    "omission": "1",
    "incorrect_location": "1",
    "incorrect_severity": "2",
    "comparison_not_present": "0",
    "omission_of_comparison": "1"
  }
}

```

```

},
"total_score": "7"

```

### Labeled data for MRscore Scoring System

```

{
  "Ground Truth Report": "Lateral view somewhat limited due to
overlying motion artifact. The lungs are low in volume. There is no
focal airspace consolidation to suggest pneumonia. A 1.2-cm
calcified granuloma just below the medial aspect of the right
hemidiaphragm is unchanged from prior study. No pleural effusions
or pulmonary edema. There is no pneumothorax. The inferior
sternotomy wire is fractured but unchanged. Surgical clips and
vascular markers in the thorax are related to prior CABG surgery.",
  "Predicted Report 0-40": {
    "Ground Truth Report": "Lateral view somewhat limited due to
overlying motion artifact. The lungs are low in volume. There is no
focal airspace consolidation to suggest pneumonia. A 1.2-cm
calcified granuloma just below the medial aspect of the right
hemidiaphragm is unchanged from prior study. No pleural effusions
or pulmonary edema. There is no pneumothorax. The inferior
sternotomy wire is fractured but unchanged. Surgical clips and
vascular markers in the thorax are related to prior CABG surgery.",
    "Predicted Report": "The patient's lungs are normal in volume
with no signs of pneumonia. There is a possibility of a small
pneumothorax. The sternotomy wire is intact and there are no
surgical clips visible.",
    "Scoring": {
      "Impression consistency": 30,
      "Impression Organ": 20,
      "Description of Lesion": 20,
      "Clinical History": 10,
      "Completeness": 10,
      "Grammar": 0,
      "Medical Terminology": 0
    },
    "Analysis": {
      "Impression consistency": "The predicted report contradicts
the ground truth report in terms of lung volume and presence of
pneumothorax.",

```

```

    "Impression Organ": "The predicted report fails to mention
the right hemidiaphragm and the thorax, which are present in the
ground truth report.",
    "Description of Lesion": "The predicted report does not
mention the 1.2-cm calcified granuloma and the fractured sternotomy
wire.",
    "Clinical History": "The predicted report does not mention
the prior CABG surgery.",
    "Completeness": "The predicted report does not include all
information from the ground truth report.",
    "Grammar": "No issues with grammar or spelling.",
    "Medical Terminology": "No issues with medical terminology."
  },
  "Final Score": 10
},
"Predicted Report 40-70": {
  "Ground Truth Report": "Lateral view somewhat limited due to
overlying motion artifact. The lungs are low in volume. There is no
focal airspace consolidation to suggest pneumonia. A 1.2-cm
calcified granuloma just below the medial aspect of the right
hemidiaphragm is unchanged from prior study. No pleural effusions
or pulmonary edema. There is no pneumothorax. The inferior
sternotomy wire is fractured but unchanged. Surgical clips and
vascular markers in the thorax are related to prior CABG surgery.",
  "Predicted Report": "Lungs are low in volume. There is no
pneumonia present but there is a 1.2-cm calcified granuloma below
the right hemidiaphragm. There are no signs of pleural effusions or
pulmonary edema. A sternotomy wire is visible but its status is
unknown. There are surgical clips present in the thorax.",
  "Scoring": {
    "Impression consistency": 30,
    "Impression Organ": 0,
    "Description of Lesion": 20,
    "Clinical History": 0,
    "Completeness": 10,
    "Grammar": 0,
    "Medical Terminology": 0
  },
  "Analysis": {
    "Impression consistency": "The predicted report does not
mention the limited lateral view due to overlying motion artifact.",

```

```
    "Impression Organ": "The predicted report correctly
identifies the organs mentioned in the ground truth report.",
    "Description of Lesion": "The predicted report does not
mention the fractured sternotomy wire.",
    "Clinical History": "The predicted report correctly mentions
the surgical clips in the thorax, indicating a history of surgery.",

    "Completeness": "The predicted report fails to include the
status of the sternotomy wire and the absence of pneumothorax.",
    "Grammar": "No issues with grammar or spelling.",
    "Medical Terminology": "No issues with medical terminology."
  },
  "Final Score": 40
},
"Predicted Report 70-100": {
  "Ground Truth Report": "Lateral view somewhat limited due to
overlying motion artifact. The lungs are low in volume. There is no
focal airspace consolidation to suggest pneumonia. A 1.2-cm
calcified granuloma just below the medial aspect of the right
hemidiaphragm is unchanged from prior study. No pleural effusions
or pulmonary edema. There is no pneumothorax. The inferior
sternotomy wire is fractured but unchanged. Surgical clips and
vascular markers in the thorax are related to prior CABG surgery.",
  "Predicted Report": "Limited lateral view due to overlying
motion artifact. Lungs are low in volume with no signs of pneumonia.
A 1.2-cm calcified granuloma is present below the right
hemidiaphragm. No pleural effusions or pulmonary edema. No
pneumothorax. The sternotomy wire is fractured but unchanged.
Surgical clips and vascular markers indicate a history of CABG
surgery.",
  "Scoring": {
    "Impression consistency": 0,
    "Impression Organ": 0,
    "Description of Lesion": 0,
    "Clinical History": 0,
    "Completeness": 0,
    "Grammar": 0,
    "Medical Terminology": 0
  },
  "Analysis": {
    "Impression consistency": "The predicted report is
consistent with the ground truth report.",
```

```
        "Impression Organ": "The predicted report correctly
identifies the organs mentioned in the ground truth report.",
        "Description of Lesion": "The predicted report correctly
describes the lesions mentioned in the ground truth report.",
        "Clinical History": "The predicted report correctly
identifies the patient's clinical history.",
        "Completeness": "The predicted report includes all
information from the ground truth report.",
        "Grammar": "No issues with grammar or spelling.",
        "Medical Terminology": "No issues with medical terminology."
    },
    "Final Score": 100
}
}
```

### C.3 Scoring Dataset Prompt Samples

#### Scoring Dataset Prompt Samples for RadCliQ Scoring System

```
{
```

"chosen": "Human: 'The ground truth report is: Impression: Tortuous aorta with prominence of ascending aortic contour. If clinical concern, could be further evaluated with chest CT. Multiple calcified granulomas. Findings: Chest PA and lateral radiograph demonstrates a tortuous aorta with questionable prominence of the ascending aortic contour. Heart size is normal. The previously noted right lower lung opacity has largely resolved with minimal residual linear opacities evident on the lateral view, likely post-inflammatory. There has been interval resolution of the previously identified right lower lung opacity. Multiple calcified nodules identified, the largest located in the left upper lung. No pleural effusion or pneumothorax evident., Assistant: 'The predicted report is: Impression: Aorta is tortuous with an unclear prominence of the ascending contour. Chest CT is advised for further assessment. Calcified granulomas are observed. Findings: The chest X-ray reveals a tortuous aorta, but the prominence of the ascending aortic contour is not evident. The heart appears enlarged, which is a new finding. The right lower lung opacity noted before has resolved, with only minor residual opacities likely due to post-inflammatory changes. Several calcified nodules are seen, with a significant one in the right upper lung. No pleural effusion or pneumothorax is present.', Human: 'Please evaluate the quality of the assistant's predicted radiology report based on the ground truth provided.'",

```

"rejected": "Human: 'The ground truth report is: Impression:
Tortuous aorta with prominence of ascending aortic contour. If
clinical concern, could be further evaluated with chest CT. Multiple
  calcified granulomas. Findings: Chest PA and lateral radiograph
demonstrates a tortuous aorta with questionable prominence of the
ascending aortic contour. Heart size is normal. The previously noted
  right lower lung opacity has largely resolved with minimal residual
  linear opacities evident on the lateral view, likely post-
inflammatory. There has been interval resolution of the previously
identified right lower lung opacity. Multiple calcified nodules
identified, the largest located in the left upper lung. No pleural
effusion or pneumothorax evident., Assistant: 'The predicted report
is: Impression: Aorta shows tortuosity with no significant
prominence of the ascending aortic contour. Suggest chest CT if
there are clinical concerns. Several calcified granulomas identified.
  Findings: Chest radiograph indicates a tortuous aorta without clear
  prominence of the ascending aortic contour. Heart size within
normal limits. Previous right lower lung opacity has mostly resolved,
  with some residual linear opacities visible on the lateral view,
suggesting post-inflammatory changes. Multiple calcified nodules
seen, predominantly in the left upper lung. No signs of pleural
effusion or pneumothorax detected.', Human: 'Please evaluate the
quality of the assistant's predicted radiology report based on the
ground truth provided.'",
"margin": [
  1,
  0,
  1,
  0,
  0,
  0,
  0,
  2
],
"chosen_score": 4,
"rejected_score": 2
}

```

### Scoring Dataset Prompt Samples for MRscore Scoring System

```
{
```

```
"chosen": "Human: 'The ground truth report is: Single frontal radiograph of the chest was performed and reveals no acute cardiopulmonary process. The cardiomediastinal and pleural structures are unremarkable. There is scarring in the upper lungs with superior traction of the hila. There is no pleural effusion or pneumothorax. Heart size is normal. Surgical hardware is seen at the right glenohumeral joint and ___ are seen within the abdomen with cardiophrenic angle may represent a small left pleural effusion as was previously seen approximately one month prior., Assistant: 'The predicted report is:Frontal chest radiograph shows no acute cardiopulmonary process. There is scarring in the upper lungs. No pleural effusion or pneumothorax. Heart size is normal.', Human: 'Please evaluate the quality of the assistant's predicted radiology report based on the ground truth provided.'",
"rejected": "Human: 'The ground truth report is:Single frontal radiograph of the chest was performed and reveals no acute cardiopulmonary process. The cardiomediastinal and pleural structures are unremarkable. There is scarring in the upper lungs with superior traction of the hila. There is no pleural effusion or pneumothorax. Heart size is normal. Surgical hardware is seen at the right glenohumeral joint and ___ are seen within the abdomen with cardiophrenic angle may represent a small left pleural effusion as was previously seen approximately one month prior., Assistant: 'The predicted report is:Frontal chest radiograph shows the heart and lungs are normal. No previous surgical hardware or abnormality is noted.', Human: 'Please evaluate the quality of the assistant's predicted radiology report based on the ground truth provided.'",
"margin": [
    0,
    20,
    0,
    0,
    0,
    0,
    0,
    0,
    20
],
"chosen_score": 40,
"rejected_score": 20
}
```



# Bibliography

- [1] R. Ambati and C. R. Dudyala. “A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering”. In: *2018 15th IEEE India Council International Conference (INDICON)*. IEEE. 2018, pp. 1–6.
- [2] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. “Palm 2 technical report”. In: *arXiv preprint arXiv:2305.10403* (2023).
- [3] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, et al. “Openflamingo: An open-source framework for training large autoregressive vision-language models”. In: *arXiv preprint arXiv:2308.01390* (2023).
- [4] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. “Qwen Technical Report”. In: *arXiv preprint arXiv:2309.16609* (2023).
- [5] S. Banerjee and A. Lavie. “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”. In: *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*. Ed. by J. Goldstein, A. Lavie, C. Lin, and C. R. Voss. Association for Computational Linguistics, 2005, pp. 65–72. URL: <https://aclanthology.org/W05-0909/>.
- [6] S. Banerjee and A. Lavie. “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”.

- In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72.
- [7] S. Bannur, K. Bouzid, D. C. Castro, A. Schwaighofer, A. Thieme, S. Bond-Taylor, M. Ilse, F. Pérez-García, V. Salvatelli, H. Sharma, et al. “Maira-2: Grounded radiology report generation”. In: *arXiv preprint arXiv:2406.04449* (2024).
- [8] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, et al. “Learning to exploit temporal structure for biomedical vision-language processing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15016–15027.
- [9] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, et al. “Making the most of text semantics to improve biomedical vision–language processing”. In: *European conference on computer vision*. Springer. 2022, pp. 1–21.
- [10] S. Bu, T. Li, Y. Yang, and Z. Dai. “Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 14194–14204.
- [11] A. Calamida, F. Nooralahzadeh, M. Rohanian, K. Fujimoto, M. Nishio, and M. Krauthammer. “Radiology-Aware Model-Based Evaluation Metric for Report Generation”. In: *arXiv preprint arXiv:2311.16764* (2023).
- [12] J. M. Z. Chaves, S.-C. Huang, Y. Xu, H. Xu, N. Usuyama, S. Zhang, F. Wang, Y. Xie, M. Khademi, Z. Yang, et al. “Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation”. In: *arXiv preprint arXiv:2403.08002* (2024).
- [13] J. M. Z. Chaves, S.-C. Huang, Y. Xu, H. Xu, N. Usuyama, S. Zhang, F. Wang, Y. Xie, M. Khademi, Z. Yang, et al. “Training small multi-modal models to bridge biomedical competency gap: A case study in radiology imaging”. In: *CoRR* (2024).

- [14] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, et al. “Meditron-70b: Scaling medical pretraining for large language models”. In: *arXiv preprint arXiv:2311.16079* (2023).
- [15] Z. Chen, Y. Zhou, A. Tran, J. Zhao, L. Wan, G. S. K. Ooi, L. T.-E. Cheng, C. H. Thng, X. Xu, Y. Liu, et al. “Medical Phrase Grounding with Region-Phrase Context Contrastive Alignment”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 371–381.
- [16] Z. Chen, Y. Shen, Y. Song, and X. Wan. “Cross-modal memory networks for radiology report generation”. In: *arXiv preprint arXiv:2204.13258* (2022).
- [17] Z. Chen, Y. Song, T.-H. Chang, and X. Wan. “Generating radiology reports via memory-driven transformer”. In: *arXiv preprint arXiv:2010.16056* (2020).
- [18] C.-H. Chiang and H.-y. Lee. “Can large language models be an alternative to human evaluations?” In: *arXiv preprint arXiv:2305.01937* (2023).
- [19] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. “Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023”. In: URL <https://lmsys.org/blog/2023-03-30-vicuna> 3.5 (2023).
- [20] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara. “Meshed-Memory Transformer for Image Captioning”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 10575–10584. DOI: [10.1109/CVPR42600.2020.01059](https://doi.org/10.1109/CVPR42600.2020.01059). URL: [https://openaccess.thecvf.com/content%5C\\_CVPR%5C\\_2020/html/Cornia%5C\\_Meshed-Memory%5C\\_Transformer%5C\\_for%5C\\_Image%5C\\_Captioning%5C\\_CVPR%5C\\_2020%5C\\_paper.html](https://openaccess.thecvf.com/content%5C_CVPR%5C_2020/html/Cornia%5C_Meshed-Memory%5C_Transformer%5C_for%5C_Image%5C_Captioning%5C_CVPR%5C_2020%5C_paper.html).
- [21] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara. “Meshed-memory transformer for image captioning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10578–10587.

- [22] D. Dai, Y. Sun, L. Dong, Y. Hao, Z. Sui, and F. Wei. "Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers". In: *arXiv preprint arXiv:2212.10559* (2022).
- [23] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. "Preparing a collection of radiology examinations for distribution and retrieval". In: *Journal of the American Medical Informatics Association* 23.2 (2016), pp. 304–310.
- [24] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li. "Transvg: End-to-end visual grounding with transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1769–1779.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).
- [26] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen. "Multiple meta-model quantifying for medical visual question answering". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 64–74.
- [27] Y. Du, Z. Fu, Q. Liu, and Y. Wang. "Visual grounding with transformers". In: *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2022, pp. 1–6.
- [28] S. Eslami, G. de Melo, and C. Meinel. "Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?" In: *arXiv preprint arXiv:2112.13906* (2021).
- [29] C. Finn, P. Abbeel, and S. Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1126–1135.
- [30] D. Ganeshan, P.-A. T. Duong, L. Probyn, L. Lenchik, T. A. McArthur, M. Retrouvey, E. H. Ghobadi, S. L. Desouches, D. Pastel, and I. R. Francis. "Structured reporting in radiology". In: *Academic radiology* 25.1 (2018), pp. 66–73.

- [31] T. M. Gemma Team, C. Hardin, R. Dadashi, S. Bhupatiraju, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, and et al. “Gemma”. In: (2024). DOI: [10.34740/KAGGLE/M/3301](https://doi.org/10.34740/KAGGLE/M/3301). URL: <https://www.kaggle.com/m/3301>.
- [32] F. M. Hall. *The Title of the Article*. <https://www.ajronline.org/doi/10.2214/ajr.175.5.1751239>. Accessed: 2023-02-24. 2000.
- [33] M. P. Hartung, I. C. Bickle, F. Gaillard, and J. P. Kanne. “How to create a great radiology report”. In: *RadioGraphics* 40.6 (2020), pp. 1658–1670.
- [34] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. “Pathvqa: 30000+ questions for medical visual question answering”. In: *arXiv preprint arXiv:2003.10286* (2020).
- [35] N. Houlsby, A. Giurghi, S. Jastrzebski, B. Morrone, Q. De Larousilhe, A. Gesmundo, M. Attariyan, and S. Gelly. “Parameter-efficient transfer learning for NLP”. In: *International conference on machine learning*. PMLR. 2019, pp. 2790–2799.
- [36] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [37] A. Huang, O. Banerjee, K. Wu, E. P. Reis, and P. Rajpurkar. “FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores”. In: *arXiv preprint arXiv:2405.20613* (2024).
- [38] W. Huang, H. Zhou, C. Li, H. Yang, J. Liu, and S. Wang. “Enhancing Representation in Radiography-Reports Foundation Model: A Granular Alignment Algorithm Using Masked Contrastive Learning”. In: *arXiv preprint arXiv:2309.05904* (2023).
- [39] Z. Huang, X. Zhang, and S. Zhang. “Kiut: Knowledge-injected u-transformer for radiology report generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19809–19818.
- [40] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, et al. “Chexpert:

- A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.
- [41] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, et al. "Radgraph: Extracting clinical entities and relations from radiology reports". In: *arXiv preprint arXiv:2106.14463* (2021).
- [42] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, et al. *Mistral 7B*. 2023. eprint: [2310.06825](https://arxiv.org/abs/2310.06825).
- [43] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen. "In defense of grid features for visual question answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10267–10276.
- [44] H. Jin, H. Che, Y. Lin, and H. Chen. "Promptmrg: Diagnosis-driven prompts for medical report generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 3. 2024, pp. 2607–2615.
- [45] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng. "MIMIC-CXR: A large publicly available database of labeled chest radiographs". In: *CoRR abs/1901.07042* (2019). arXiv: [1901.07042](https://arxiv.org/abs/1901.07042). URL: <http://arxiv.org/abs/1901.07042>.
- [46] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". In: *arXiv preprint arXiv:1901.07042* (2019).
- [47] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. "Mdetr-modulated detection for end-to-end multi-modal understanding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1780–1790.
- [48] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. "Referitgame: Referring to objects in photographs of natural scenes". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 787–798.

- [49] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar. "MMBERT: multimodal BERT pretraining for improved medical VQA". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1033–1036.
- [50] J.-H. Kim, J. Jun, and B.-T. Zhang. "Bilinear attention networks". In: *Advances in neural information processing systems* 31 (2018).
- [51] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. "A dataset of clinically generated visual questions and answers about radiology images". In: *Scientific data* 5.1 (2018), pp. 1–10.
- [52] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day". In: *arXiv preprint arXiv:2306.00890* (2023).
- [53] J. Li, D. Li, S. Savarese, and S. Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *arXiv preprint arXiv:2301.12597* (2023).
- [54] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang. "Dynamic graph enhanced contrastive learning for chest x-ray report generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3334–3343.
- [55] M. Li and L. Sigal. "Referring transformer: A one-step approach to multi-task visual grounding". In: *Advances in neural information processing systems* 34 (2021), pp. 19652–19664.
- [56] Y. Li, Y. Liu, Z. Wang, X. Liang, L. Liu, L. Wang, L. Cui, Z. Tu, L. Wang, and L. Zhou. "A comprehensive study of GPT-4V's multimodal capabilities in medical imaging". In: *medRxiv* (2023), pp. 2023–11.
- [57] Y. Li, Z. Wang, Y. Liu, L. Wang, L. Liu, and L. Zhou. "KARGEN: Knowledge-enhanced Automated Radiology Report Generation Using Large Language Models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 382–392.
- [58] Y. Li, X. Liang, Z. Hu, and E. P. Xing. "Hybrid retrieval-generation reinforced agent for medical image report generation". In: *Advances in neural information processing systems* 31 (2018).

- [59] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, and Y. T. Lee. *Textbooks Are All You Need II: phi-1.5 technical report*. 2023. arXiv: [2309.05463](https://arxiv.org/abs/2309.05463) [cs.CL].
- [60] C.-Y. 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. URL: <https://aclanthology.org/W04-1013>.
- [61] C.-Y. Lin. “Rouge: A package for automatic evaluation of summaries”. In: *Text summarization branches out*. 2004, pp. 74–81.
- [62] A. Liu, Y. Guo, J. Yong, and F. Xu. “Multi-grained radiology report generation with sentence-level image-language contrastive learning”. In: *IEEE Transactions on Medical Imaging* (2024).
- [63] A. Liu, Y. Guo, J.-h. Yong, and F. Xu. “Multi-grained Radiology Report Generation with Sentence-level Image-language Contrastive Learning”. In: *IEEE Transactions on Medical Imaging* (2024).
- [64] C. Liu, Y. Tian, W. Chen, Y. Song, and Y. Zhang. “Bootstrapping Large Language Models for Radiology Report Generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 17. 2024, pp. 18635–18643.
- [65] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou. “Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 13753–13762. DOI: [10.1109/CVPR46437.2021.01354](https://doi.org/10.1109/CVPR46437.2021.01354). URL: [https://openaccess.thecvf.com/content/CVPR2021/html/Liu%5C\\_Exploring%5C\\_and%5C\\_Distilling%5C\\_Posterior%5C\\_and%5C\\_Prior%5C\\_Knowledge%5C\\_for%5C\\_Radiology%5C\\_Report%5C\\_CVPR%5C\\_2021%5C\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Liu%5C_Exploring%5C_and%5C_Distilling%5C_Posterior%5C_and%5C_Prior%5C_Knowledge%5C_for%5C_Radiology%5C_Report%5C_CVPR%5C_2021%5C_paper.html).
- [66] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou. “Exploring and distilling posterior and prior knowledge for radiology report generation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 13753–13762.

- [67] Y. Liu, Y. Li, Z. Wang, X. Liang, L. Liu, L. Wang, L. Cui, Z. Tu, L. Wang, and L. Zhou. "A systematic evaluation of gpt-4v's multi-modal capability for chest x-ray image analysis". In: *Meta-Radiology* (2024), p. 100099.
- [68] Y. Liu, Z. Wang, Y. Li, X. Liang, L. Liu, L. Wang, and L. Zhou. "MRScore: Evaluating Radiology Report Generation with LLM-based Reward System". In: *arXiv preprint arXiv:2404.17778* (2024).
- [69] Y. Liu, Z. Wang, D. Xu, and L. Zhou. "Q2atransformer: Improving medical vqa via an answer querying decoder". In: *International Conference on Information Processing in Medical Imaging*. Springer, 2023, pp. 445–456.
- [70] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.
- [71] J. Lu, C. Xiong, D. Parikh, and R. Socher. "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 3242–3250. DOI: [10.1109/CVPR.2017.345](https://doi.org/10.1109/CVPR.2017.345). URL: <https://doi.org/10.1109/CVPR.2017.345>.
- [72] J. Lu, C. Xiong, D. Parikh, and R. Socher. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 375–383.
- [73] A. Lukaszewicz, J. Uricchio, and G. Gerasymchuk. "The art of the radiology report: practical and stylistic guidelines for perfecting the conveyance of imaging findings". In: *Canadian Association of Radiologists Journal* 67.4 (2016), pp. 318–321.
- [74] Meta. *Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date*. <https://ai.meta.com/blog/meta-llama-3/>. Accessed: 2024-05-20. 2024.

- [75] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran. "Overcoming data limitation in medical visual question answering". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 522–530.
- [76] A. Nicolson, J. Dowling, and B. Koopman. "Improving chest X-ray report generation by leveraging warm starting". In: *Artificial intelligence in medicine* 144 (2023), p. 102633.
- [77] OpenAI. "GPT-4 Technical Report". In: *ArXiv abs/2303.08774* (2023). URL: <https://api.semanticscholar.org/CorpusID:257532815>.
- [78] S. Ostmeier, J. Xu, Z. Chen, M. Varma, L. Blankemeier, C. Bluethgen, A. E. Michalson, M. Moseley, C. Langlotz, A. S. Chaudhari, et al. "GREEN: Generative Radiology Report Evaluation and Error Notation". In: *arXiv preprint arXiv:2405.03595* (2024).
- [79] M. Pahadia, S. Khurana, H. Geha, and S. T. I. Deahl. "Radiology report writing skills: A linguistic and technical guide for early-career oral and maxillofacial radiologists". In: *Imaging Science in Dentistry* 50.3 (2020), p. 269.
- [80] Y. Pan, T. Yao, Y. Li, and T. Mei. "X-Linear Attention Networks for Image Captioning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 10968–10977. DOI: [10.1109/CVPR42600.2020.01098](https://doi.org/10.1109/CVPR42600.2020.01098). URL: [https://openaccess.thecvf.com/content%5C\\_CVPR%5C\\_2020/html/Pan%5C\\_X-Linear%5C\\_Attention%5C\\_Networks%5C\\_for%5C\\_Image%5C\\_Captioning%5C\\_CVPR%5C\\_2020%5C\\_paper.html](https://openaccess.thecvf.com/content%5C_CVPR%5C_2020/html/Pan%5C_X-Linear%5C_Attention%5C_Networks%5C_for%5C_Image%5C_Captioning%5C_CVPR%5C_2020%5C_paper.html).
- [81] K. Papineni, S. Roukos, T. Ward, and W. Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pp. 311–318. DOI: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135). URL: <https://aclanthology.org/P02-1040/>.
- [82] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

- [83] PEIR digital library. <http://peir.path.uab.edu/library/index.php?/category/2>.
- [84] C. Pellegrini, M. Keicher, E. Özsoy, and N. Navab. “Rad-ReStruct: A Novel VQA Benchmark and Method for Structured Radiology Reporting”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 409–419.
- [85] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, and M. Keicher. “Rad-Dialog: A large vision-language model for radiology report generation and conversational assistance”. In: *arXiv preprint arXiv:2311.18681* (2023).
- [86] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. “Kosmos-2: Grounding Multimodal Large Language Models to the World”. In: *arXiv preprint arXiv:2306.14824* (2023).
- [87] C. C. Radiology. *Why Clinical History Is Essential for Diagnoses*. <https://radiologyblog.cincinnatichildrens.org/why-clinical-history-essential-for-diagnoses/>. Accessed: 2023-02-24. 2017.
- [88] N. Reimers and I. Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. 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 2019, Hong Kong, China, November 3-7, 2019*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Association for Computational Linguistics, 2019, pp. 3980–3990. DOI: [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410). URL: <https://doi.org/10.18653/v1/D19-1410>.
- [89] F. Ren and Y. Zhou. “Cgmvqa: A new classification and generative model for medical visual question answering”. In: *IEEE Access* 8 (2020), pp. 50626–50636.
- [90] Y. Shi, D. Peng, W. Liao, Z. Lin, X. Chen, C. Liu, Y. Zhang, and L. Jin. “Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation”. In: *arXiv preprint arXiv:2310.16809* (2023).
- [91] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al. “Towards expert-level medical question answering with large language models”. In: *arXiv preprint arXiv:2305.09617* (2023).

- [92] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. “CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT”. In: *arXiv preprint arXiv:2004.09167* (2020).
- [93] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. “Learning to summarize with human feedback”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021.
- [94] J. Sun, D. Wei, Z. Xu, D. Lu, H. Liu, L. Wang, and Y. Zheng. “You’ve Got Two Teachers: Co-evolutionary Image and Report Distillation for Semi-supervised Anatomical Abnormality Detection in Chest X-Ray”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 363–373.
- [95] Z. Sun, M. Lin, Q. Zhu, Q. Xie, F. Wang, Z. Lu, and Y. Peng. “A scoping review on multimodal deep learning in biomedical images and texts”. In: *Journal of Biomedical Informatics* (2023), p. 104482.
- [96] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).
- [97] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. “Multimodal few-shot learning with frozen language models”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 200–212.
- [98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [99] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett. 2017, pp. 5998–6008. URL: <https://proceedings>.

- [neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html).
- [100] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. “Cider: Consensus-based image description evaluation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.
- [101] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. “Show and tell: A neural image caption generator”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3156–3164. DOI: [10.1109/CVPR.2015.7298935](https://doi.org/10.1109/CVPR.2015.7298935). URL: <https://doi.org/10.1109/CVPR.2015.7298935>.
- [102] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. “Self-consistency improves chain of thought reasoning in language models”. In: *arXiv preprint arXiv:2203.11171* (2022).
- [103] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou. “Automated radiographic report generation purely on transformer: A multicriteria supervised approach”. In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2803–2813.
- [104] Z. Wang, L. Liu, L. Wang, and L. Zhou. “Metransformer: Radiology report generation by transformer with multiple learnable expert tokens”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 11558–11567.
- [105] Z. Wang, L. Liu, L. Wang, and L. Zhou. “R2gengpt: Radiology report generation with frozen llms”. In: *Meta-Radiology* 1.3 (2023), p. 100033.
- [106] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou. “A medical semantic-assisted transformer for radiographic report generation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 655–664.
- [107] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. “Emergent abilities of large language models”. In: *arXiv preprint arXiv:2206.07682* (2022).

- [108] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.
- [109] J. R. Wilcox. "The written radiology report." In: *Applied Radiology* 35.7 (2006).
- [110] C. Wu, J. Lei, Q. Zheng, W. Zhao, W. Lin, X. Zhang, X. Zhou, Z. Zhao, Y. Zhang, Y. Wang, et al. "Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis". In: *arXiv preprint arXiv:2310.09909* (2023).
- [111] C. Wu, J. Liu, X. Wang, and R. Li. "Differential networks for visual question answering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 8997–9004.
- [112] Y. Wu, S. Wang, H. Yang, T. Zheng, H. Zhang, Y. Zhao, and B. Qin. "An early evaluation of gpt-4v (ision)". In: *arXiv preprint arXiv:2310.16534* (2023).
- [113] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by F. R. Bach and D. M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, 2015, pp. 2048–2057. URL: <http://proceedings.mlr.press/v37/xuc15.html>.
- [114] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, et al. "Qwen2 Technical Report". In: *CoRR* (2024).
- [115] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao. "Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V". In: *arXiv preprint arXiv:2310.11441* (2023).
- [116] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao. "Knowledge Matters: Radiology Report Generation with General and Specific Knowledge". In: *Medical Image Analysis* (2021).
- [117] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang. "The dawn of lmms: Preliminary explorations with gpt-4v (ision)". In: *arXiv preprint arXiv:2309.17421* (2023).

- [118] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. "Stacked attention networks for image question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 21–29.
- [119] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. "mplug-owl: Modularization empowers large language models with multimodality". In: *arXiv preprint arXiv:2304.14178* (2023).
- [120] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, et al. "Evaluating progress in automatic chest x-ray radiology report generation". In: *Patterns* 4.9 (2023).
- [121] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, et al. "Evaluating progress in automatic chest x-ray radiology report generation". In: *Patterns* 4.9 (2023).
- [122] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. Fonseca, H. Lee, Z. Shakeri, A. Ng, et al. *Radiology Report Expert Evaluation (ReXVal) Dataset*. 2023.
- [123] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675* (2019).
- [124] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "BERTScore: Evaluating Text Generation with BERT". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- [125] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu. "When Radiology Report Generation Meets Knowledge Graph". In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2020).
- [126] W. Zhao, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. "Ratescore: A metric for radiology report generation". In: *arXiv preprint arXiv:2406.16845* (2024).

- 
- [127] Y. Zhao, Z. Lin, D. Zhou, Z. Huang, J. Feng, and B. Kang. “Bubogpt: Enabling visual grounding in multi-modal llms”. In: *arXiv preprint arXiv:2307.08581* (2023).
- [128] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji. “Seqtr: A simple yet universal network for visual grounding”. In: *European Conference on Computer Vision*. Springer, 2022, pp. 598–615.
- [129] D. M. Ziegler, N. Stiennon, J. Wu, T. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. “Fine-tuning language models from human preferences”. In: *arXiv preprint arXiv:1909.08593* (2019).