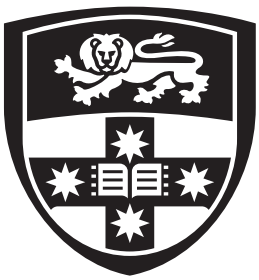
Multi-agent System and Reinforcement Learning in Medical Report Generation

Pengyu Wang



THE UNIVERSITY OF
SYDNEY

A thesis submitted in fulfillment of the requirements
of the degree of Master of Philosophy

Faculty of Engineering
School of Computer Science
The University of Sydney

April 2026

Statement of Originality

This is to certify that, to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

During the preparation of this thesis, ChatGPT was used for the purposes of text enhancement, including improving sentence structure, clarifying phrasing, and paraphrasing content for better readability. Where any text was modified by generative AI, the author then reviewed the resulting content for any errors, inaccuracies or biases, and modified it as required.

The author takes full responsibility for the submitted thesis and ensures the work is their own and has used generative AI within the parameters of use, see University of Sydney generative AI guide for researchers .

Authorship Attribution Statement

Chapter 3 of this thesis is based on a publication P. Wang, S. Ye, U. Naseem, and J. Kim "MRGAgents: A Multi-Agent Framework for Improved Medical Report Generation with Med-LVLMs." in 2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2025, accepted. I designed the method, conducted the experiments, analysed the data and drafted the manuscript.

Chapter 4 of this thesis is based on a publication P. Wang, S. Ye, U. Naseem, and J. Kim "MRG-R1: Reinforcement Learning for Clinically Aligned Medical Report Generation" in preparation. I designed the method, conducted the experiments, analysed the data and drafted the manuscript.

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.

Abstract

Medical large vision-language models (Med-LVLM) have enabled automatic medical report generation (MRG), yet two fundamental challenges persist which undermines diagnostic utility. First, models frequently exhibit a normality bias, under-detecting abnormalities and results in incomplete coverage of clinically relevant findings across disease categories. Second, mainstream training relies on token-level imitation, capturing lexical style rather than clinical correctness and report-level semantic alignment. To address these issues, this thesis presents a new framework comprising two complementary pillars: (i) MRGAgents, a multi-agent system (MAS) that introduces disease-specific multi-agent modeling to decompose reporting into condition-focused subtasks for more balanced and comprehensive coverage and, (ii) MRG-R1, a finetuning-paradigm via semantic-driven reinforcement learning (SRL) that directly optimizes report-level clinical correctness and factual alignment beyond surface fluency.

The proposed MRGAgents comprises specialized agents trained on curated disease-specific subsets of IU X-ray and MIMIC-CXR, endowing each agent with stronger discrimination and descriptive capability for its target conditions. At inference, their outputs are aggregated to achieve a better balance between normal and abnormal findings and to ensure more comprehensive category-specific descriptions pertinent to diagnosis. Empirically, MRGAgents consistently improved coverage and abnormality reporting over state-of-the-art baselines and contemporary counterparts, mitigating missed findings.

From the training method perspective, MRG-R1 introduces SRL with Group Relative Policy Optimization (GRPO) and a margin Chexbert cosine similarity (MCCS)

reward computed on key radiologic findings. This directly optimizes report-level clinical-label agreement and semantic consistency beyond surface-level style mimicry. In this project, Medical Report Generation with SRL (MRG-R1) was evaluated on two datasets: IU X-Ray and MIMIC-CXR using clinical efficacy (CE) metrics. MRG-R1 achieved state-of-the art performance with CE-F1. Ablation studies identify two primary drivers of improvement. First, the proposed MCCS reward provided finer-grained supervision compared to traditional CE-F1-based objectives, enabling the model to capture more nuanced clinical semantics. Second, incorporating an explicit reasoning-to-report process encouraged structured generation, leading to more coherent and diagnostically accurate reports. Notably, these gains were achieved with minimal computational overhead, maintaining practicality for real-world deployment.

Overall, the architectural (multi-agent, disease-specific modeling) and training method (clinically aligned reinforcement learning) contributions advance report comprehensiveness, abnormality sensitivity, and clinical correctness for chest X-ray report generation.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Jinman Kim, for his guidance, support, and encouragement throughout my MPhil. Professor Kim not only drew on his broad expertise and rich research experience to help me navigate technical challenges; more importantly, he shaped the way I think about research itself. His mentorship has been a compass on my path from novice to more mature researcher. Every meeting with him brought clarity and inspiration. His patient explanations, precise feedback, and steady counsel made it possible for me to complete the MPhil smoothly. Beyond academia, Professor Kim also offered invaluable advice on life and career planning. I am profoundly grateful for the time and effort he invested in me.

I am deeply thankful to my parents for their unwavering support during my MPhil. They shouldered all of my expenses and, when I encountered setbacks in my research, they provided unconditional emotional support. Their steadfast love has been my anchor.

My sincere thanks also go to my associate supervisors, Professor Usman Naseem and Professor Adam G. Dunn. They offered generous, unreserved guidance, keen insights into emerging issues, and timely, practical feedback—especially on framing research questions and shaping the writing of papers. Their contributions strengthened my work in countless ways.

I am indebted to my collaborators, Shuchang Ye and Yongpei Ma. Under their guidance, I completed my first research project and wrote my first paper during the MPhil. Their sharp intuition for frontiers of research and their exemplary engineering rigor helped me avoid many detours. I am truly grateful for their hands-on advice both in experimentation and in writing.

I also wish to thank friends and colleagues in the BDAV lab who supported me during my MPhil: Boyuan Tan, Guangcheng Andrew Zhang, Hoijoon Jung, Hao Wang, Huang Yi, Jiadi Dong, Jiyong Wang, Professor Lei Bi, Mingxiao Tu, Xiao Wang, Xumou Zhang, Yaoyao Yue, Dr. Yuan Yuan, and Yupeng Zhang.

My heartfelt thanks to my roommates, Luoyuan Zhang and Zhenhuan Zhang, for their companionship and understanding.

I am grateful to the University of Sydney for the outstanding training and support over the past year: abundant opportunities for academic communication, high-quality experimental facilities, and a clean, well-equipped working environment.

I also thank the many organizations that provide open datasets and the open-source communities: GitHub, Hugging Face, and ModelScope for making research more accessible. My work benefited greatly from the exceptional software and tools created by the developer communities behind Python, Anaconda, CUDA, PyTorch, Transformers, Weights & Biases (wandb), DeepSpeed, PEFT, and FlashAttention. I further acknowledge the model research and engineering teams and the organizations behind them who open-sourced Qwen, Llama, DeepSeek, and others. Finally, I would like to thank the Stanford Machine Learning Group for its contributions to the chest X-ray research community.

List of Publications

Published or Accepted:

1. P. Wang, S. Ye, U. Naseem, and J. Kim "MRGAgents: A Multi-Agent Framework for Improved Medical Report Generation with Med-LVLMs." in 2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2025

Under Preparation, Review or Revision:

1. P. Wang, S. Ye, U. Naseem, and J. Kim "MRG-R1: Reinforcement Learning for Clinically Aligned Medical Report Generation"
(Journal Paper, In Preparation)

2. Y. Ma, Z. Duan, P. Wang, U. Naseem, A. Dunn, and J. Kim "Ask Twice, Look Twice: Training-Free Consistency Filtering for Reliable Medical VQA"
(Scientific Report, under review)

Contents

List of Figures

List of Tables

List of Acronyms

Med-LVLM	Medical Large Vision Language Model
SRL	Semantic-driven Reinforcement Learning
GRPO	Group Relative Policy Optimization
MCCS	Margin Chexbert Cosine Similarity
MRG	Medical Report Generation
LVLM	Large vision-language model
MAS	Multi-Agent Systems
PPO	Proximal Policy Optimization
CoT	Chain-of-Thought
CXR	Chest-X-ray
AI	Artificial Intelligence
LSTM	Long Short-Term Memory
VLP	Vision-Language Pretraining
LLM	Large Language Model
EHR	Electronic Health Record
SAM	Segment Anything
IE	Information Extraction
RL	Reinforcement Learning
TF-IDF	Term Frequency–Inverse Document Frequency
SFT	Supervised Finetuning
CT	Computed Tomography
MRI	Magnetic Resonance Imaging

Chapter 1

Introduction

1.1 Background and Motivation

Medical imaging is central to modern diagnostics, and radiology reports are the primary vehicle for communicating evidence-based findings and guiding treatment. However, producing detailed and accurate reports is time-consuming and labor-intensive, with mounting imaging volumes exacerbating workload, turnaround times, and cognitive burden [30, 81, 104]. Interpreting large image sets and translating observations into narratives that cover both normal and abnormal findings demands sustained expertise, and under pressure even experienced radiologists can miss subtle abnormalities, misspecify negations or uncertainties, and apply terminology inconsistently across cases [30, 65]. Against this backdrop, automatic medical report generation (MRG) has emerged as a promising means to alleviate workload, improve efficiency and consistency, and reduce human error so clinicians can focus on complex cases and patient care [81, 104]. Recent research has advanced MRG from early encoder-decoder models to more sophisticated Transformer-based frameworks that better map complex visual inputs into coherent textual descriptions and integrate multi-modal information [9, 11]. Moreover, large vision-language models (LVLM) tailored to medicine (e.g., BioMedGPT, LLaVA-Med) have demonstrated encouraging fluency and adaptability across downstream tasks, including report generation [37, 106]. While LVLM-based

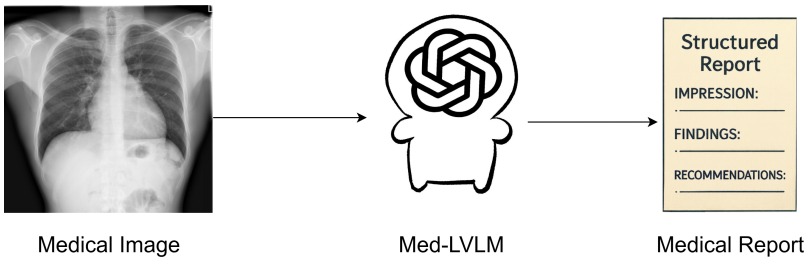


Figure 1.1 – A conceptual framework of medical report generation

report generators have improved in fluency, however monolithic models remain vulnerable to normality bias and incomplete coverage of clinically relevant regions. This is primary because heterogeneous and long-tailed chest X-ray pathologies are averaged within a single model's representation [9, 11, 19, 54]. In such settings, because most training reports describe normal findings while only a small fraction contain abnormalities, the model's learning process becomes dominated by the abundant "normal" patterns. As a result, abnormal cases contribute only weak and noisy gradients, leading the model to under-represent disease-related features in its learned representation space and reducing its sensitivity to individual findings. To address such challenges in other complex domains, multi-agent systems (MAS) have been widely explored for their ability to leverage task specialization and collaborative reasoning. By decomposing a problem into coordinated subtasks handled by specialized agents, MAS approaches have demonstrated improvements in both efficiency and decision quality across diverse fields, including healthcare [31, 79, 98]. Given that medical report generation similarly involves multiple interdependent subtasks: detecting, describing, and contextualizing findings, this paradigm of structured specialization and collaboration offers a natural direction for mitigating the limitations of monolithic models.

Although modern LVLM pipelines have improved fluency and stylistic coherence, their primary training objective remains token-level imitation which is predicting the next word based on local context. This process rewards lexical resemblance to reference reports rather than ensuring report-level clinical accuracy, meaning the model learns to mimic the surface form of text instead of understanding or verifying underlying medical facts. As a result, the generated outputs often take the form of grammatically fluent yet semantically shallow narratives, which may not fully align

with image evidence and tend to exhibit fragility in expressing negation or uncertainty [53, 65, 68]. To overcome this limitation, several studies have attempted to inject semantic supervision through methods such as contrastive learning and image-text matching, multitask objectives with auxiliary classification or localization heads, and dynamic traceback curricula. While these approaches enhance alignment between image and text to some extent, they remain insufficiently fine-grained: they often overlook subtle findings and entity-relation structures [28, 40, 92, 94, 95, 103, 109]. Consequently, existing methods still focus on stylistic imitation rather than optimizing for clinical faithfulness and report-level semantic correctness.

Despite steady advances in Med-LVLM pipelines, current MRG remains misaligned with clinical needs in two fundamental ways. First, data and task imbalance combined with cross-disease heterogeneity lead to a persistent normality bias and incomplete coverage: normal patterns dominate corpora while minority disease signals receive weak gradients. A single monolithic Med-LVLM further averages across heterogeneous conditions, diluting disease-specific cues and under-calling abnormalities, resulting in reports that downplay or omit clinically significant findings and regions [9, 11, 19, 54]. Second, the prevailing learning objective is token-level imitation (next-token likelihood, n-gram overlap), which optimizes surface fluency and stylistic resemblance rather than report-level clinical correctness; this mismatch produces plausible but unsupported statements and fragile handling of negation or uncertainty [53, 65, 68]. Motivated by these gaps, this thesis focuses on the problem of automatically generating radiology reports from chest X-ray (CXR) images that are faithful to image evidence and useful at the point of care. The objective is to produce reports that (1) deliver complete and sensitivity-aware coverage of clinically relevant findings, capturing both normal and abnormal observations without omissions or defaulting to "normal". (2) Maintain report-level clinical correctness, including consistent treatment of negation and uncertainty.

1.2 Contribution

To address these challenges, this thesis presents two complementary contributions that advance both architectural design and training paradigms to produce higher-quality, clinically aligned medical reports in the CXR domain. To rigorously evaluate these contributions, the framework is applied to chest X-ray data—chosen for its clinical ubiquity, standardized reporting conventions, and the availability of large, publicly benchmarked datasets (e.g., MIMIC-CXR and IU X-Ray), making it an ideal and representative testbed for developing and assessing medical report generation systems. First contribution delivers the design and implementation of a disease-specialized multi-agent LVLM architecture (MRGAgents) that decomposes reporting into CheXbert-anchored, disease-specific sentence tasks and aggregates multiple agent outputs into a structured narrative. In the second contribution, the development of a semantic-driven reinforcement learning framework (MRG-R1) is introduced that directly optimizes report-level clinical correctness via Group Relative Policy Optimization (GRPO)—a value-free, groupwise policy-gradient method that stabilizes training by comparing samples within a batch rather than relying on a learned critic—combined with clinically informed rewards and instruction-guided formatting. Together, these contributions address normality bias, improve coverage of clinically relevant regions and findings, and align generation with clinical semantics within the chest-X-ray MRG domain.

1. **MRGAgents: A disease-specialized, multi-agent LVLM architecture for MRG applications.**

   Radiology reporting is inherently disease-centric: narratives are organized around findings, and radiologists reason and communicate through condition-specific descriptions. Building on this observation, the MRGAgents framework introduces a disease-specialized multi-agent large vision-language model (LVLM) architecture that reformulates report generation as a set of condition-specific, sentence-level subproblems anchored by CheXbert supervision. Each agent is

trained on a curated, disease-specific subset to strengthen its ability to discriminate and describe its target findings, and their parallel outputs are subsequently integrated into a coherent, structured report. This architecture leverages task specialization and coordinated aggregation to counter normality bias, enhance sensitivity to abnormal findings, and improve completeness across clinically relevant regions and categories. The resulting framework produces balanced and clinically consistent chest X-ray narratives, addressing the core coverage inherent in monolithic Med-LVLM approaches.

2. **MRG-R1: A semantic-driven reinforcement learning framework that directly optimizes report-level clinical correctness.**

   The MRG-R1 framework introduces a semantic-driven reinforcement learning paradigm designed to optimize report-level clinical correctness beyond token-level imitation. This approach formulates post-training as a reinforcement learning process that directly maximizes agreement between generated reports and key radiologic findings, grounding the reward signal in CheXbert-derived clinical labels to emphasize factual completeness and semantic alignment rather than lexical overlap. To make this optimization stable and computationally efficient for large vision-language models (LVLM), MRG-R1 employs GRPO, groupwise variant related to Proximal Policy Optimization (PPO), which stabilizes learning through case-conditioned relative advantages instead of a learned critic, enabling reliable updates under sparse and non-differentiable rewards [76, 78]. A margin-based cosine reward further refines this process by measuring semantic consistency between predicted and reference findings. Complementing this optimization, lightweight instruction-guided formatting elicits an interpretable "reasoning $\rightarrow$ report" structure without annotated Chain-of-Thought (CoT), improving factual consistency, the handling of negation and uncertainty, and the overall diagnostic utility of generated chest X-ray reports.

1.3 Thesis Organization

The remainder of this thesis is organized as follows:

- Chapter 2 reviews related work on medical report generation and vision-language modeling, with emphasis on chest X-ray applications. It surveys Transformer-based and LVLM approaches, multi-agent methodologies in clinical AI, and semantic supervision strategies, highlighting gaps that motivate architectural specialization and clinically aligned optimization.

- Chapter 3 presents MRGAgents, a disease-specialized multi-agent framework for medical report generation. The chapter details the CheXbert-anchored, sentence-level task decomposition, the curation of disease-specific training subsets from IU X-ray and MIMIC-CXR, the parallel inference and structured aggregation mechanism, and the evaluation protocol centered on coverage of clinically relevant findings and balanced normal/abnormal reporting.

- Chapter 4 introduces MRG-R1, a semantic-driven reinforcement learning framework for clinically aligned report generation. It describes the GRPO-based optimization scheme, the margin-based cosine reward derived from CheXbert labels to capture report-level clinical semantics, and an instruction-guided formatting strategy to elicit explicit reasoning, together forming a training procedure that aligns generation with clinical objectives on chest X-ray data.

- Chapter 5 first summarizes the overall work presented in this thesis, and then explores its strengths, limitations, and potential future directions in this field.

Chapter 2

Literature Review

2.1 Background and Chapter Overview

Medical report generation (MRG) converts medical images into radiology-style reports that document clinically relevant findings, impressions, and recommendations in a clinician-familiar format [50, 91]. Interest in MRG has grown because radiology workloads and reporting pressures continue to rise, where variability, omissions, and clinical uncertainty affect report quality and turnaround time [6, 65, 72]; by automating parts of the image-to-text pipeline, AI systems can help improve efficiency, consistency, and clinical fidelity. Recent surveys synthesize this trend, summarizing multimodal pipelines (images, clinical text, knowledge), structured-report generation, public datasets, evaluation methods, and the remaining barriers to practical [25, 63, 91].

This chapter first surveys advances in MRG in Section 2.2; then Medical Large Vision-Language Models in Section 2.3; followed by Multi-Agent Systems in Healthcare in Section 2.4; Semantic Supervision in Section 2.5; Reinforcement Learning for Large Vision-Language Models in Section 2.6.

2.2 Medical Report Generation

MRG aims to produce sectioned narratives that capture anatomy, attributes, and clinically meaningful qualifiers within radiology reports. Beyond linguistic fluency, MRG models must ensure report-level factual adequacy, comprehensive coverage of clinically salient findings, and internal consistency. Historically, MRG emerged from the image-captioning paradigm [60], where encoder–decoder architectures were trained via maximum-likelihood objectives to map visual representations to natural-language descriptions; a canonical example is Show and Tell [89], which couples a deep visual feature extractor with a recurrent sequence generator to produce captions. Early MRG work adapted this recipe to longer, multi-sentence outputs with clinical terminology. Jing et al. [28] introduced hierarchical LSTMs with co-/cross-attention with auxiliary disease tags to improve sentence planning and grounding for chest-X-ray findings. Retrieval–generation hybrids such as HRGR-Agent [41] coupled a generator with a retrieval policy to stabilize long-form outputs, while TieNet [92] jointly learned imag-text embeddings for thorax disease classification and produced preliminary reports, tightening visual-textual coupling. With Transformers, MRG shifted beyond recurrent decoders. R2Gen [11] introduced a memory-driven Transformer that caches global cues to manage long-range context across sentences and performs strongly on IU X-Ray and MIMIC-CXR. Beyond R2Gen, R2GenCMN [9] augments the memory-driven Transformer with cross-modal memory under teacher-forcing training, achieving stronger maximum-likelihood baselines on IU X-Ray and MIMIC-CXR. Knowledge-aware variants (e.g., PPKED [44]) mitigate visual/textual bias but keep the same learning target. Despite architectural differences, most systems still optimize token-level likelihood, favoring lexico-syntactic overlap over clinically grounded semantics.

Recently, Med-LVLMs have pushed capacity and generality. LLaVA-Med [37] instruction-tunes a vision–language backbone on biomedical image-text pairs for dialogue, captioning, and Visual Question Answering (VQA); Med-Flamingo [55] adapts Open-Flamingo [1] for few-shot generative medical VQA with physician blind review; HuatuoGPT-

Vision [8] injects medical visual knowledge at scale into Qwen2-VL [90] and Qwen2.5VL [2] using an LLaVA-style training pipeline; MedGemma [77] adapts Gemma [86] to the medical domain via instruction tuning and evaluated on biomedical tasks. CheXagent [12] targets chest-X-ray interpretation and multi-task evaluation via a curated instruction datasets. Radiology-focused variants such as CXR-LLaVA [35] tailor LLaVA [46] to chest X-rays and study zero-/few-shot reporting or recognition; broader generalist biomedical models (e.g., BioMedGPT [106]) and radiology foundation efforts (e.g., RadFM [97]) pursue unified pretraining across modalities. Despite these advances, recent benchmarking still finds that general-purpose LLM/LVLM pipelines underperform on clinically faithful reporting without explicit clinical objectives or specialized training [93].

Beyond lexical overlap, recent work evaluates clinical factuality using finding-aware labelers and structured graphs. In particular, Metrics such as RadGraph-F1 [26] and the RadCliQ [105] correlate better with radiologists' error judgments than BLEU [64]/ROUGE [42]/BERTScore [107] , and are now recommended for CXR report generation. This shift underscores the need to optimize clinical content rather than style alone.

Chest-Xray corpora such as MIMIC-CXR [29], CheXpert [24], PadChest [7], VinDr-CXR [57] underpin most progress; automatic labelers (e.g., CheXpert [24]/CheXbert [80], NegBio [65]) enable scalable supervision yet introduce polarity and uncertainty noise, especially for devices and localization.

Despite richer prompting, retrieval, and tool integration, most medical report generators—from memory-driven Transformers (e.g., R2Gen, R2GenCMN) to instruction-tuned LVLMs (e.g., LLaVA-Med, HuatuoGPT-Vision)—remain limited in two fundamental ways. First, monolithic architectures trained on long-tailed, heterogeneous datasets tend to overfit to dominant "normal" patterns, leading to normality bias and incomplete coverage of clinically significant abnormalities. Second, their learning objective is still dominated by token-level maximum likelihood estimation (MLE), which optimizes surface fluency and lexical resemblance rather than clinically grounded semantics, producing text that may read well but lacks diagnostic fidelity.

2.3 Medical Large Vision-Language Model

Med-LVLMs adapt the LLM-plus-visual-encoder paradigm to clinical images for instruction following, report-like generation, and multimodal reasoning. Early "generalist biomedical" LVLMs demonstrated breadth across many medical tasks: BioMedGPT [106] proposed an open, lightweight generalist LVLM trained over diverse biomedical tasks; LLaVA-Med [47] showed cost-efficient instruction tuning from general LLaVA [37] to biomedical use; Med-Flamingo [55] extended OpenFlamingo for [1] few-shot generative medical VQA with physician evaluation; and Med-PaLM Multimodal (Med-PaLM M) [87] pursued a single model operating over text, medical images, and even genomics, with radiologist side-by-side preferences on chest-X-ray reporting. Recent general-purpose medical LVLMs further broaden this landscape: MedGemma [77] introduces a family of Gemma-3 based [86] medical vision-language foundation models aimed at broad clinical tasks; Lingshu [99] presents a medical-specialized MLLM trained with a multi-stage curriculum and a unified evaluation toolkit (MedEvalKit) for multimodal and textual benchmarks. These efforts highlight the promise and the difficulty of clinical grounding at scale.

Radiology-centric foundation models have emerged to specialize capacity and data curation. RadFM [97] targets a unified radiology foundation model with large multimodal datasets (MedMD) and reports strong results across 2D/3D scans. Meanwhile, CXR-LLaVA [35] fine-tunes LLaVA [47] for chest X-rays; CheXagent [12] builds a chest-X-ray foundation model by using CheXinstruct/CheXbench resources and reports prospective time-savings in clinician editing; Collectively, these radiology-focused LVLMs pursue reporting assistance and VQA while probing hallucination control, uncertainty handling, and domain adaptation.

Beyond radiography, pathology-focused LVLMs address gigapixel slides and multi-resolution context. PathAsst [82] pairs a pathology-tuned encoder with Vicuna-13B [110] and a pathology toolkit for generative assistance; PA-LLaVA [15] combines a pathology image-text encoder [73] with scale-invariant connectors before instruction tuning for whole-slide VQA. These works illustrate that medical LVLMs often require

bespoke data pipelines and connectors even within a shared architecture paradigm.

A critical enabler across many medical LVLMs is vision-language pretraining (VLP) on paired image-report corpora and biomedical figure-caption collections. ConVIRT [108] pioneered contrastive pretraining from chest X-ray and report pairs; GLoRIA [23] added global–local alignment between image regions and report tokens; BioViL/BioViL-T [5] introduced temporal CXR-text pretraining with progression signals; large-scale figure–caption models such as PMC-CLIP and BiomedCLIP (PMC-15M) [43] provide broad biomedical backbones that many LVLMs later adapt or build upon. These VLP advances improve data efficiency, zero/few-shot transfer, and phrase-level grounding—key ingredients for clinically faithful generation.

Recent trends emphasize multilingual and data-centric scaling. HuatuoGPT-Vision curated and denoised PubMedVision ($\approx$ 1.3M medical VQA samples) to inject medical visual knowledge into multimodal LLMs [8]; Qilin-Med-VL [48] targeted a Chinese medical LVLM with two-stage alignment and a million-scale image–text dataset to broaden language coverage.

Despite steady progress, current medical LVLMs still lack mechanisms for fine-grained clinical grounding and disease-specific reasoning, leaving gaps between fluent generation and diagnostic reliability.

2.4 Multi-Agent System in Healthcare

Healthcare multi-agent systems (MAS) orchestrate cooperating LLM/LVLM agents with distinct roles (retrieval, reasoning, tool use, auditing) to address the sequential, uncertain, and multimodal nature of clinical work. Recent frameworks include MDAgents [31], which automatically assigns collaboration topologies to teams of LLMs for complex medical decision-making; ArgMed-Agents [21], which promotes explainable clinical reasoning via self-argumentation and conflict graphs, making deliberations auditable. These systems aim to reduce single-model brittleness and expose intermediate rationale for safety review.

Microsoft's Sequential Diagnosis with Language Models [59] casts diagnosis as a stepwise encounter, building the Sequential Diagnosis Benchmark (from 304 NEJM-CPC cases) where an agent must iteratively query for new findings before updating the hypotheses: an agentic setup that better reflects real clinics than static QA. Contemporary analyses and coverage also describe orchestration among multiple models for differential diagnosis under cost constraints. Complementary testbeds such as AgentClinic [75] simulates doctor–patient–tool interactions across specialties and languages, and shows substantial performance drops when models must plan, gather evidence, and handle bias. Emerging benchmarks like MedAgentsBench [84] further stress multi-step reasoning, diagnosis formulation, and treatment planning to differentiate thinking models and agent frameworks.

In structured data settings, EHRAgent [79] equips an LLM with code-execution and memory to decompose multi-table electronic health record (EHR) questions into executable tool chains, achieving few-shot complex query answering; in multi-modal tool ecosystems, MMedAgent [36] learns to select among specialist medical tools (spanning modalities and tasks), showing how a coordinator agent can improve performance while remaining extensible as new tools are added. These works demonstrate MAS as a natural fit for clinical data orchestration, where different competencies, query planning, coding, image analysis, are delegated to specialised agents or tools.

Despite rapid progress in clinical reasoning, diagnosis, and decision-support, current medical multi-agent systems have rarely been applied to medical report generation (MRG). Existing frameworks focus primarily on diagnostic dialogue, retrieval, and multimodal reasoning rather than narrative synthesis. Extending the MAS paradigm to MRG remains an open opportunity, where disease-specialized agents could collaboratively produce comprehensive, clinically faithful reports aligned with radiologic workflows.

2.5 Semantic Supervision

To mitigate the mismatch between token-level objectives and clinical goals, recent work augments MRG with semantic supervision that directly encodes medical knowledge, grounding, and clinical consistency. Knowledge-centric methods explicitly encode medical structure. KERP [38] learns an abnormality graph and paraphrases it into text to improve entity–relation fidelity; GSKET [102] integrates general (graph-based) and specific (retrieved case) knowledge to guide sentence planning. Knowledge graph and knowledge base variants continue this line, for example, Attributed Abnormality Graph (ATAG) [100] and Dynamic Graph Enhanced Contrastive Learning [39] that update graph structure and add contrastive/matching losses for finer semantics.

A complementary thread strengthens image-text grounding via matching or contrastive signals. Co-training a generator with image-text matching (ITM) heads ("self-boosting" [95]) improves clinical alignment by penalizing mismatched pairs. Reinforced Cross-modal Alignment [66] introduces an RL objective over a cross-modal memory to better couple visual and textual cues; and segment-enhanced contrastive learning (MSCL [109]) leverages segmentation (e.g., SAM [32]) to focus alignment on clinically meaningful regions of interest and reduce dataset bias. Retrieval-assisted systems operationalize this alignment at inference time: X-REM [27] learns a contrastive matching score to retrieve report sentences conditioned on the image, improving grounding and reducing unsupported statements. CXRMate [58] introduces longitudinal semantic rewards that leverage follow-up consistency signals to reduce hallucinations and improve clinically coherent reporting in chest X-rays.

Moving closer to clinically grounded objectives [45], labeler- and information extraction (IE) -based supervision turns clinical extractors into training signals. Clinically Accurate Chest X-ray Report Generation [45] optimizes an RL reward tied to clinical coherence; CheXbert [80] provides a BERT-based automatic labeler widely used to score presence/absence of 14 observations. On the IE side, RadGraph [26] introduces chest-x-ray entity–relation annotations, enabling RadGraph-based rewards that directly optimize factual completeness/correctness. Beyond rewards, Dynamic Trace-

back Learning [103] regularizes causal consistency by masking/back-tracing across modalities so tokens can be "explained" by visual evidence, reducing spurious correlations.

Despite these advances, most existing methods incorporate semantic or clinical signals only as auxiliary objectives rather than core optimization targets. As a result, token-level likelihoods still dominate the training signal, leaving a gap between linguistic fluency and clinically faithful reasoning—calling for reinforcement learning frameworks that directly optimize report-level clinical correctness.

2.6 Reinforcement Learning for Large Vision-Language Model

Post-training for LLMs and LVLMs increasingly adopts preference-based objectives that align models with human or AI feedback [3, 62]. The standard Reinforcement Learning from Human Feedback (RLHF) framework [62] first trains a reward model on human preference data and then optimizes the policy using Proximal Policy Optimization (PPO) [76] under a Kullback–Leibler (KL) divergence constraint [33]. This setup, popularized by InstructGPT, uses PPO's clipped objective to ensure stable updates during finetuning. To reduce dependence on costly human annotation, Reinforcement Learning from AI Feedback (RLAIF) or Constitutional AI [3] replaces human preferences with AI-generated critiques guided by rules or principles. Direct Preference Optimization (DPO) [67] simplifies RLHF by eliminating the separate reward model and directly fitting model outputs to preference pairs. Group Relative Policy Optimisation (GRPO) [78] further improves efficiency by comparing multiple sampled responses within a group to compute relative advantages, avoiding learnt value functions while improving reasoning performance. Other recent variants, such as Odds Ratio Preference Optimization (ORPO) [20] and Kahneman-Tversky Optimization (KTO) [18], explore lighter-weight formulations that remove the reference model or use binary desirability signals inspired by human decision theory.

In computer vision and captioning, reinforcement learning has long been used to bridge the gap between likelihood training and evaluation metrics. MIXER [69] gradually shifts from teacher-forced learning to sampled decoding and applies REINFORCE [96] to optimize non-differentiable metrics such as BLEU [64] and ROUGE [42]. Self-Critical Sequence Training (SCST) [71] stabilizes learning by using the model's greedy caption as a baseline, achieving strong results on CIDEr [88] and BLEU. Later work introduced SPIDEr (SPICE + CIDEr) [49] and other actor–critic or embedding-based methods to better align captions with visual semantics. More recently, the reward functions themselves have evolved: CLIP-based [13] rewards emphasize recognizability and distinctiveness, while PACScore [74] offers a learned, human-correlated evaluation model. At the same time, RL-free methods have appeared—such as DiCO [56], which distills CLIP/PAC preferences directly into the model, and DMO [83], which performs offline reward-weighted augmentation—offering compute-efficient alternatives that retain semantic fidelity without full reinforcement learning loops.

Despite these advances, most preference-based post-training frameworks still optimize linguistic quality rather than domain-grounded semantics. In clinical report generation, where factual and diagnostic precision are critical, this creates a gap between fluent text generation and clinically faithful reasoning—highlighting the need for clinically informed, semantic-driven reinforcement learning.

Chapter 3

MRGAgents: A Multi-Agent Framework for Improved Medical Report Generation with Med-LVLMs

3.1 Introduction

Medical imaging plays a critical role in clinical diagnostics, with radiology reports serving as essential tools for communicating diagnostic findings and guiding treatment decisions. However, the process of generating detailed and accurate medical reports remains time-consuming and labor-intensive [81, 104]. Automated medical report generation has emerged as a promising solution, offering the potential to reduce radiologists' workload, improve efficiency, and minimize human errors, and thus allowing clinicians to focus more on complex cases and patient care. Over the past few years, significant research efforts have been dedicated to medical report generation (MRG), with models progressively advancing towards more sophisticated Transformer-based frameworks. These models aim to map complex visual data into coherent textual descriptions, improving the integration of multi-modal information.

Recent advancements [9, 37, 106] in Large Vision-Language Models (LVLMs) have accelerated the progress in the field, enabling models to process and understand both medical images and textual data more effectively. Notable examples, such as BioMedGPT [106] and LLaVA-Med [37], have demonstrated their versatility across various medical downstream tasks, particularly in MRG [9, 11]. Despite these advancements, existing methods continue to face significant challenges. First, prior studies [19] have shown that current models tend to be biased toward normal findings. Because normal descriptions dominate most training corpora, the model gradients are driven mainly by these majority patterns; consequently, within a single optimization step, the network receives only weak or noisy signals for minority diseases, and critical abnormalities are often under-represented in the learned feature space. This imbalance translates into reports that downplay or altogether omit rare but clinically significant findings, and thus limiting their diagnostic usefulness. Second, reports generated by Medical LVLMs (Med-LVLMs) often lack comprehensiveness [54], which is defined as the ability of a generated medical report to cover all clinically relevant findings, including both normal and abnormal observations, without omitting key diagnostic information.

Meanwhile, recent studies have shown that multi-agent systems (MAS)—collections of cooperating autonomous agents— can enhance task efficiency, improve decision-making [31], and optimise resource allocation across diverse domains, including healthcare [79], autonomous driving, and large-scale data processing [98]. By leveraging the principle of task specialization, multi-agent systems can decompose complex problems into sub-tasks. This lets specialized agents focus on distinct aspects of the task, thereby improving both efficiency and accuracy [31, 85]. However, existing medical multi-agent frameworks primarily focus on Med-QA (Medical Question Answering) [85] and Med-VQA(Medical Visual Question Answering) [21, 31, 36], and their applications to medical report generation remains unexplored. Given that MRG is inherently a complex task requiring fine-grained descriptions across multiple disease categories, a multi-agent framework is proposed to decompose the report-generation process into more manageable sub-problems.

In this study, **M**edical **R**eport **G**eneration **Agents** (**MRGAgents**) is introduced as the **first multi-agent LVLM-based framework for medical report generation**. Building on recent advances in multi-agent systems [21, 31, 36, 85], the approach leverages the capacity of multi-agent collaboration to meet the unique demands of medical report generation. The pipeline begins with CheXbert [80], a BERT-based classifier that assigns each sentence in a chest X-ray report to one or more of 14 predefined observations (e.g., cardiomegaly, pleural effusion, pneumothorax) and labels them as positive, negative, or uncertain. CheXbert outputs are then used to partition each report into sentence-level, disease-specific subsets, which subsequently serve as training data for 13 specialized agents, each corresponding to a specific disease category. Radiology reports are naturally organised around findings such as heart size, lung opacity, or pleural effusion, and clinicians review prior studies by the same disease labels. Training sentences at this granularity therefore produces output that matches how radiologists read, dictate, and archive reports. At inference, a chest X-ray is broadcast to all agents, which run in parallel, each agent returns one concise statement that explicitly indicates whether its own finding is present, absent, or uncertain. Concatenating the 13 sentences yields a structured report that systematically covers every disease category, potentially delivering higher recall and richer detail than a single Med-LVLM.

The key contributions are as follows:

- The first multi-agent LVLM framework for medical report generation (MRG) is presented, in which 13 task-specific agents operate in parallel—each producing a disease-focused sentence, and the resulting sentences are aggregated into a single, clinically meaningful report.
- A sentence-level task decomposition strategy is introduced to segment reports into disease-specific components, enabling finetuning of multiple specialized agents on distinct medical conditions. Furthermore, disease-specific subsets are curated from the IU X-ray and MIMIC-CXR datasets, demonstrating that MRGAgents achieves improved performance across most disease categories.

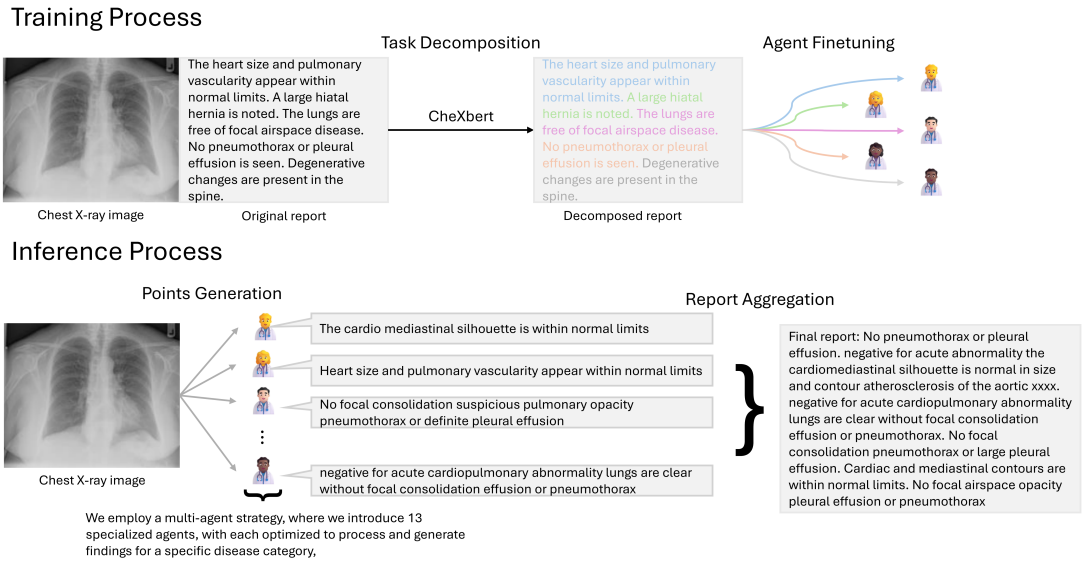


Figure 3.1 – Overall Framework of the proposed MRGAgents.

- Experimental results on two widely used medical report generation benchmark datasets demonstrate superior performance of MRGAgent compared to existing Med-LVLM baselines.

3.2 Methods

3.2.1 Overall Framework of MRGAgents

The MRGAgents framework (Figure 3.1) operates in two phases: task decomposition and category-specific finetuning.

Training phase. Instead of finetuning a single Med-LVLM on all findings, CheXbert [80] is first applied to divide each report into 13 disease-specific sentence sets. Thirteen lightweight agents are then fine-tuned, each using only the sentences that describe its assigned disease. Sentences with no positive finding are labelled "No Finding" and omitted from training. This specialisation allows every agent to focus on one pathology, improving both precision and coverage.

Inference phase. Given a chest X-ray, each agent generates a single sentence for its disease category. These sentences are finally concatenated to produce the complete radiology report.

3.2.2 Task Decomposition: Sentence-Level Report Splitting

Medical reports typically comprise several sentences, each conveying different types of clinical information—normal anatomy, abnormal findings, lesion location, and so on. Prior work [37, 106] on automatic report generation often produces overly general statements that miss such fine-grained details. To address this limitation, a sentence-level task-splitting strategy is adopted to enhance both the specificity and completeness of the generated reports.

CheXbert [80], a BERT-based model for multi-label classification of radiology text, is employed to assign each sentence to one or more of 14 predefined observations: Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices, and No Finding. For every observation it predicts a label of positive, negative, uncertain, or blank.

These fine-grained labels allow us to build dedicated training subsets for specialised agents. By separating each report into disease-specific components, every agent learns to generate precise, clinically relevant sentences for its assigned category. Aggregating the outputs of all agents therefore yields a final report that is markedly more detailed and informative than those produced by single Med-LVLM.

3.2.3 Agent Finetuning and Inference Process

Each agent in the framework is fine-tuned with the original BioMedGPT hyperparameter settings [106]: AdamW optimiser, learning-rate 5×10^{-5}, batch size 16, weight decay 0.01, a linear warm-up over the first 10% of three training epochs, and early stopping on the validation loss. Following the BioMedGPT fine-tuning pipeline,

agent training adopts a standard autoregressive sequence-to-sequence maximum-likelihood objective. Given the input chest X-ray (encoded as visual tokens) together with the instruction/prompt text x, the model generates the target disease-specific sentence $y = (y_1, \ldots, y_T)$ token-by-token in a left-to-right manner under teacher forcing, and the parameters are optimized by minimizing the negative log-likelihood (token-level cross-entropy):

$$\mathcal{L}_{\mathrm{MLE}}(\theta) = -\sum_{t=1}^{T} \log p_\theta \left(y_t \mid y_{<t}, x\right). \tag{3.1}$$

At inference time, every agent generates one finding sentence for its assigned disease category, and these sentences are concatenated to form the final radiology report. Both training stage and inference stage are deployed on a server with i7-5930K CPU and two 48GB NVIDIA A6000 GPUs. MRGAgents does not introduce a new or disease-specific prompt; all agents follow the default BioMedGPT captioning/report-generation instruction. Agent specialization is achieved purely through disease-wise data partitioning. The exact prompt setting is provided in Appendix A.1.1.

3.3 Experiment

3.3.1 Datasets

The proposed framework was evaluated on two benchmark medical report generation datasets: IU X-ray [17] and MIMIC-CXR [29].

Throughout the rest of this paper, the following two-letter abbreviations are employed: EC = Enlarged Cardiomediastinum, CM = Cardiomegaly, LO = Lung Opacity, LL = Lung Lesion, ED = Edema, CS = Consolidation, PN = Pneumonia, AT = Atelectasis, PT = Pneumothorax, PF = Pleural Effusion, PO = Pleural Other, FX = Fracture, SD = Support Devices, and NF = No Finding.

IU X-ray includes 7,470 chest X-ray images paired with 3,955 reports, while MIMIC-CXR contains 377,110 images and 277,835 reports. Both datasets follow their official splits, with IU X-ray divided into 4,118 training, 588 validation, and 2,764 testing

Table 3.1 – Sentence distribution of IU X-ray and MIMIC-CXR.

	IU X-ray			MIMIC-CXR		
	training	validation	test	training	validation	test
EC	1546	244	452	156,405	1213	1394
CM	2486	378	720	136,645	1093	2107
LO	2660	362	762	102,548	752	2306
LL	310	50	100	14,702	162	313
ED	1704	228	460	75,091	587	1367
CS	242	32	70	109,220	879	1260
PN	1396	222	416	32,083	248	613
AT	162	26	50	61,139	441	1074
PT	176	20	38	198,670	1557	2441
PF	3044	436	860	232,670	1850	3296
PO	3180	456	904	5,635	47	167
FX	38	4	8	20,810	131	324
SD	202	28	50	103,934	786	2507
NF	230	38	68	929,135	7168	12,149

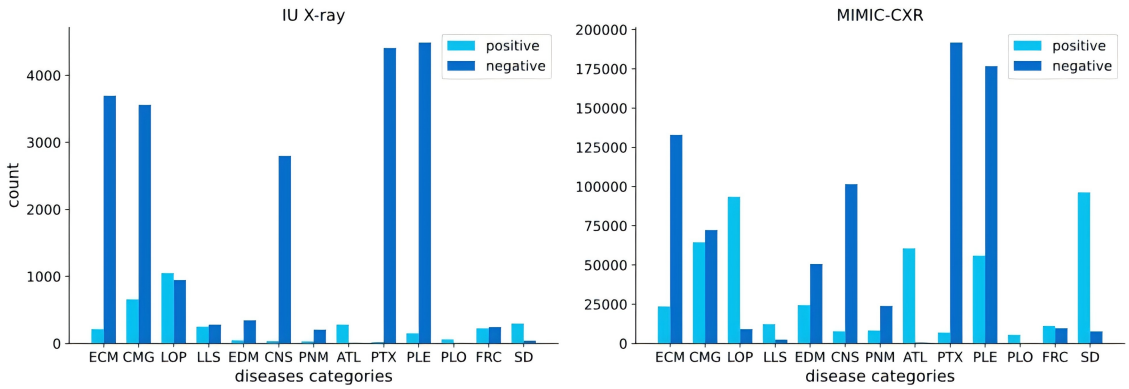


Figure 3.2 – The distribution of positive and negative sentences in each disease.

samples, and MIMIC-CXR split into 270,790 training, 2,130 validation, and 3,858 testing samples.

Each report is segmented into individual sentences and automatically labelled with CheXbert [80]. Only sentences marked as positive or negative for at least one observation are retained for finetuning the disease-specific agents. If a sentence mentions multiple categories, it is supplied to every corresponding agent. The resulting sentence-level label statistics are summarised in Table 3.1. To provide a more intuitive view of category prevalence, Figure 3.3 plots a bar chart whose x-axis lists the 13 disease categories and whose y-axis reports the number of labelled sentences.

3.3.2 Experimental Design

The performance of the proposed MRGAgents framework was benchmarked against established models for medical report generation to demonstrate its effectiveness. BioMedGPT is used as the base model for MRGAgents framework. Each agent is trained on sentences corresponding to a specific disease category, In particular, no agent is trained for the "no finding" category. All agents are fine-tuned based on the default configuration of the BioMedGPT finetuning script. Additionally, accuracy was computed for each disease to assess the ability of MRGAgents to correctly identify medical conditions.

3.3.3 Evaluation metrics

Three standard natural language generation (NLG) metrics are reported: ROUGE-L [42], METEOR [4], and CIDEr [88]. ROUGE-L measures the longest common subsequence between the generated and reference reports, while METEOR aligns tokens through both exact and synonym matching. CIDEr computes a TF-IDF–weighted n-gram similarity that rewards clinical details appearing across multiple references. For clinical efficacy evaluation, the protocol of Chen et al. [9] is followed: each generated report is converted into 13 binary disease labels using CheXbert, and macro-averaged precision, recall, and F1 scores are calculated against the ground-truth labels.

3.4 Result and Discussion

3.4.1 Comparison with Existing Models

Table 3.2 presents the comparison results between proposed MRGAgents and the state-of-the-art MRG methods, for examples, BioMedGPT and R2Gen. Compared to the single-agent BioMedGPT, MRGAgents resulted in gains on the IU X-ray benchmark. For language quality, the CIDEr score increased from 0.401 to 0.426, and METEOR improved from 0.191 to 0.205. The higher scores shows that splitting the report into disease-specific descrption sentences help the system capture more of phrasing and detail found. On the clinical side, precision metric improved from 0.360 to 0.369, and recall also improved from 0.354 to 0.376, indicating its enhanced ability to capture clinical findings.

The increased performance is further evident with on the larger MIMIC-CXR dataset. METEOR results almost doubled from 0.072 to 0.144, and ROUGE-L increased from 0.144 to 0.178. Clinical recall shows the most improvement from 0.314 to 0.382, making it the second-highest among all models, which is particularly important for reducing missed diagnoses. Precision also increased from 0.290 to 0.369. Although the resulting macro-averaged F1 (0.258) was lower than BioMedGPT's 0.286, recall

Table 3.2 – Comparisons of MRGAgents with state-of-the-art MRG methods. '*'indicates that the results were obtained through reproduction. The best results are highlighted in bold. Methods with additional annotations are showed in gray for reference. In the table, the BioMedGPT results on IU X-ray are reported as follows: the NLG metrics are taken directly from the original paper, while the CE metrics are obtained using the authors' released checkpoint. For MIMIC-CXR, all metrics are reproduced implementation following the settings described in the original work.

Data	Model	CE Metrics			NLG Metrics		
		Precision	Recall	F1	MT	CDr	RG-L
IU	R2GenCMN [9]	-	-	-	0.191	-	0.375
	R2Gen [11]				0.142		0.371
	LKAM [101]	-	-	-	-	0.407	**0.399**
	BioMedGPT* [106]	0.360	0.354	**0.355**	-	0.401	-
	MGSK [102]	-	-	-	-	0.382	0.381
	MRGAgents	**0.369**	**0.376**	0.346	**0.205**	**0.426**	0.331
MIMIC	R2GenCMN [9]	0.334	0.275	0.278	0.142	-	**0.278**
	R2Gen [11]	0.333	0.273	0.276	0.142		0.277
	LKAM [101]	**0.420**	0.339	**0.352**	-	**0.111**	0.274
	BioMedGPT* [106]	0.290	0.314	0.286	0.072	0.02	0.144
	MGSK [102]	0.458	0.348	0.371	-	0.203	0.284
	MRGAgents	0.369	**0.382**	0.258	**0.144**	0.031	0.178

improvement was considered particularly valuable in clinical practice, as it reduced the risk of missed abnormalities.

When compared to other state-of-the-art models, MRGAgents remained competitive. On IU X-ray, it achieved the highest CIDEr (0.426) and METEOR (0.205) scores among all evaluated models, while maintaining strong CE metrics. In the MIMIC-CXR dataset, MGSK achieved the highest CE metrics, likely due to its external knowledge source, which supplements training data with additional clinical information. The lower ROUGE-L score may be due to the increased length and structural differences in the generated reports, which, despite being more comprehensive, reduce the longest common subsequence overlap with the reference reports. LKAM indeed achieves the highest ROUGE-L on IU X-ray (0.399) and the best CE precision/F1 on MIMIC-CXR (0.420/0.352), reflecting its advantage in exploiting structured knowledge signals and cross-modal alignment to improve clinical correctness. In contrast,

MRGAgents focuses on multi-agent specialization, which tends to improve coverage and detail via category-specific reasoning, explaining why it leads on several NLG metrics while remaining competitive on CE metrics.

These results demonstrate that MRGAgents enables each model to pick up subtle, category-specific cues that a single Med-LVLM system tends to overlook, resulting in reports that are both more detailed and more faithful to the reference findings than those produced by the original BioMedGPT.

3.4.2 Quantitative Analysis in Disease Level

Beyond whole-report metrics, the evaluation also examines how effectively each model recognizes individual pathologies. For this experiment, BioMedGPT was trained as a single, generic decoder, whereas MRGAgents employed 13 disease-specific agents produced by the task-decomposition strategy described in Section 3. BioMedGPT is selected as a baseline because it represents a strong, general-purpose biomedical LVLM trained across multiple medical tasks, providing a fair reference for evaluating the benefits of disease-specific specialization. Table 3.3 summarises the per-category accuracies.

On IU X-ray, MRGAgents either matched or improved BioMedGPT in 11 out of 13 categories. The highest gain occurred with the previously under-detected disease categories: "Consolidation" increased from 0.171 to 0.857 and "Pleural Effusion" from 0.288 to 0.979. It further detected conditions that BioMedGPT missed entirely, such as "Lung Lesion" (0 to 0.480) and "Pneumonia" (0 to 0.404).

A similar improvement is observed on the much larger MIMIC-CXR benchmark, reinforcing the trend consistently seen across all evaluations conducted: across datasets and metrics, MRGAgents consistently outperforms its baseline model. MRGAgents outperformed the baseline in 10 categories, with increases for "Cardiomegaly" (0.036 to 0.758), "Atelectasis" (0.103 to 0.987) and "Pleural Effusion" (0.636 to 0.772). It also recovered "Pleural Other" (0 to 0.904) and "Fracture" (0 to 0.185), two classes that BioMedGPT failed to recognise. The only clear shortfall is the "Support Devices"

class, whose highly variable appearance remains challenging for the visual pipeline. In MIMIC-CXR, the Support Devices category covers a heterogeneous set of objects, such as pacemaker or ICD leads, central venous and PICC lines, chest tubes, endotracheal or tracheostomy tubes, and nasogastric/orogastric tubes, whose shapes, sizes, and positions vary substantially across cases [24]. Beyond this category-specific weakness, a deeper inspection of model outputs reveals several recurrent failure modes that are not fully captured by aggregate report-level metrics. First, omission of low-prevalence or visually subtle findings remains a common error pattern, especially for categories with limited training sentences (e.g., LL and FX) or weak visual salience. Although MRGAgents substantially improves coverage compared with BioMedGPT, some abnormalities are still missed when the visual evidence is faint, overlapped by surrounding structures, or described with highly variable reporting styles. Second, support-device under-recognition and device-type confusion persists because the current disease-specialized decomposition is optimized primarily for pathology findings, rather than for elongated, small, and procedurally placed devices. Third, label-consistent but text-structure-misaligned generation can occur, where the generated report contains clinically relevant statements but differs from the reference in wording order, granularity, or sentence composition. This helps explain why MRGAgents improves recall and several NLG metrics while not always improving ROUGE-L, which is sensitive to sequence overlap.

These failure modes have distinct clinical implications. Accordingly, MRGAgents is positioned as a report drafting assistant rather than an autonomous reporting system; in real-world use, all generated reports must be reviewed, edited, and signed off by a qualified clinician before being used for clinical decision-making. Missed abnormal findings primarily affect diagnostic sensitivity and may delay further investigation or treatment escalation, whereas support-device errors may affect procedural safety and post-procedural monitoring, particularly in acute-care settings where chest radiographs are routinely used to confirm device placement and detect device-related complications. By contrast, wording or structural mismatch without factual error mainly affects report standardization and automatic metric scores, rather than immediate

Table 3.3 – Comparison of the proposed MRGAgents with previous studies on the curated subset of IU X-ray, evaluating the accuracy of disease classification. The best results are highlighted in bold.

	IU X-ray		MIMIC-CXR	
	BioMedGPT	MRGAgents	BioMedGPT	MRGAgents
EC	0	**0.04**	0.374	**0.641**
CM	**0.856**	0.850	0.036	**0.758**
LO	0.076	**0.163**	0.005	**0.481**
LL	0	**0.480**	0	**0.556**
ED	0.035	**0.583**	0.199	**0.715**
CS	0.171	**0.857**	0.818	**0.834**
PN	0	**0.404**	0	**0.261**
AT	0	**0.160**	0.103	**0.987**
PT	0	**0.053**	0.885	**0.962**
PF	0.288	**0.979**	0.636	**0.772**
PO	0	0	0	**0.904**
FX	0	0	0	**0.185**
SD	0	**0.440**	**0.115**	0

clinical decision-making. Overall, these results indicate that report decomposition not only improves recall but also broadens disease coverage, particularly for subtle or infrequent findings. For example, lung lesion (LL) has only 310 training sentences in IU X-ray, yet its accuracy climbs from 0 to 0.480. Fracture (FX) has just 38 training sentences, but it still rises from 0 to 0.185 on MIMIC-CXR. Likewise, Pleural Other (PO), with fewer than 6,000 sentences in the MIMIC-CXR training split, improves from 0 to 0.904. These cases confirm that decomposing the task by disease not only boosts overall recall but also broadens coverage to subtle or infrequent findings that a single-decoder model tends to miss. Although running 13 independent agents introduces a modest increase in computational load, the additional latency is imperceptible to radiologists and is offset by reductions in missed abnormalities, follow-up corrections, and double-reading overhead. These observations also suggest several directions for future improvement, including a dedicated device-focused agent (or support-device pathway), stronger visual grounding for small elongated objects, and additional factuality-oriented constraints to reduce omissions while preserving the recall gains of task decomposition.

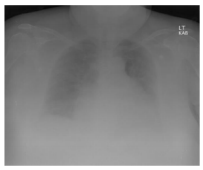
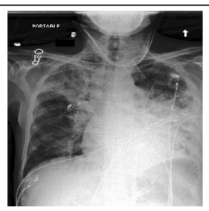
Image	Ground Truth	BioMedGPT	MRGAGENTS
 IU X-ray /CXR2285_IM-0870	XXXX sternotomy XXXX and mediastinal postsurgical changes. Stable cardiomegaly. Crowded bronchovascular and interstitial markings XXXX related to low lung volumes and technique. Grossly stable appearance of the lungs compared to prior exam without overt edema or gross airspace consolidation.	1 No acute cardiopulmonary abnormality 2 Cardiomegaly with marked tortuosity of the thoracic aorta The lungs are clear	Cardiac and mediastinal contours are within normal limits. Heart size is enlarged. low lung volumes with bibasilar opacities. cardiomegaly with central pulmonary vascular prominence and coarsened interstitial markings concerning for interstitial edema. No focal consolidation pneumothorax or large pleural effusion. No pneumonia effusions edema pneumothorax adenopathy nodules or masses. lung volumes are low with central bronchovascular crowding and patchy basilar atelectasis. No pneumothorax or pleural effusion. No pleural effusion or pneumothorax. bibasilar airspace disease left greater than right with small effusions. No acute bony abnormality. cardiomegaly and mild degenerative changes of the thoracic spine. right subclavian central line is noted with tip overlying the right brachiocephalic vein.
 MIMIC-CXR/ 43a707e7-f762b3ea-4b033501-64409bec-21b98cef	Again seen is a left PICC in the upper to mid SVC. \n Innumerable metastatic pulmonary nodules are present. There are continued\n multifocal hazy opacities, with confluent consolidation in the left lower\n lobe. Right upper lobe collapse is unchanged. Moderate left and small right\n pleural effusions. moderate cardiomegaly, and central venous congestion\n persist. No pneumothorax.	Right internal jugular central venous catheter tip terminates in the right atrium unchanged in position compared to the prior study	Heart size is enlarged. There is a left lower lobe opacity in the right lung base. The mediastinal contours are normal. Nodular opacities projecting over the right lung base are compatible with known metastatic disease. The patient is status post median sternotomy and aortic valve replacement. There is no pleural effusion or pneumothorax. There is no pneumothorax. bilateral pleural effusions with associated atelectasis are unchanged. Right apical pleural thickening is unchanged. No displaced fracture is identified. There is mild pulmonary edema. There is no focal consolidation effusion or pneumothorax. There is no focal consolidation concerning for pneumonia

Figure 3.3 – Examples of generated reports, with different text colors highlighting various medical descriptions for comparison with the Ground Truth. All reports generated by MRGAgents consist of 13 sentences, each corresponding to a specific disease category.

3.4.3 MRGAgents Visualization

Figure 3.3 presents a case study comparing reports generated by MRGAgents, BioMedGPT, and the ground truth, where MRGAgents produced most comprehensive and clinically relevant reports compared to the baseline models. In the first example, MRGAgents accurately identified key clinical findings such as "Cardiomegaly", "Pleural Effusion", and "Low Lung Volumes", which were absent with BioMedGPT. This improvement likely stems from the agent-based design, which systematically addresses all relevant disease categories.

MRGAgents has several limitations, as it fails to describe support devices present in the images. This highlights the need for further refinement to improve coverage of non-disease-related findings, such as medical implants and assistive devices.

3.5 Conclusions

In summary, MRGAgents is introduced as the first multi-agent, Med-LVLM-based framework for radiology report generation. By splitting the task into 13 disease-specific sub-agents—each fine-tuned on sentences that mention its target pathology—we move beyond the limitations of monolithic decoders and produce reports that are both richer in clinical detail and more diagnostically accurate. Extensive tests on IU X-ray and MIMIC-CXR show consistent gains over a strong BioMedGPT baseline: higher CIDEr and METEOR scores, markedly better recall of critical findings, and only a negligible increase in inference latency thanks to parallel execution. These results demonstrate that targeted specialisation is a practical way to boost report quality without onerous computational costs, and they open the door to future extensions such as device-focused agents, ensemble distillation, and knowledge-base integration.

Chapter 4

MRG-R1: Reinforcement Learning for Clinically Aligned Medical Report Generation

4.1 Introduction

Automatic medical report generation (MRG) produces radiology-style narratives from medical images, documenting clinically relevant findings, impressions, and recommendations in a format familiar to clinicians. This is a timely and important technology, as clinicians face mounting challenges with the growing volume of imaging studies due to the increasing clinical use of medical imaging [34]. Efficiently interpreting large image sets and translating observations into comprehensive diagnostic reports requires significant expertise and time, motivating efforts to automate this process. Even for experienced radiologists, the process is labor-intensive and vulnerable to error under workload pressure [30], missing subtle abnormalities or uncertainties mis-specified [65], and terminology applied inconsistently across cases. At the same time, the adoption of more complex imaging workflows—such as high-resolution cross-sectional scans, dynamic/multi-phase protocols, and multi-modality fusion—further elevates interpretation demands and error risk. Automatic medical report generation has

therefore emerged as a promising approach to alleviating radiologists' workload, improving efficiency, and enhancing consistency and clinical fidelity.

Recently, automatic medical report generation has garnered significant interest and achieved substantial advancements with deep learning (subset of AI) applied and optimized to healthcare applications [10, 11, 28, 92, 103]. Accurately capturing the full semantics of a radiology report is critical for clinical utility, as diagnostic reasoning depends on consistent representation of findings, attributes, and anatomical context across the entire narrative. Yet most existing methods are trained with token-level supervision, which optimizes next-word prediction and surface n-gram overlap. Consequently, these models often produce linguistically fluent but clinically inaccurate reports, as token-level losses fail to enforce report-level semantic consistency. Early methods for MRG adopted encoder-decoder architecture with CNN backbones to extract visual features and LSTM/GRU decoders producing sentences [28, 41, 92]. With the advent of Transformers, models improved global context modeling and parallelization, enabling stronger cross-modal alignment and better handling of long-range dependencies across multiple sentences and report sections [10, 44]. Most recently, large vision-language models (LVLMs)—which couples high-capacity visual encoders with instruction-tuned language models—have shown impressive zero-/few-shot fluency and style transfer, and can be augmented with retrieval or tool use [35, 103, 106]. Despite these advances, the standard training paradigm remains misaligned with clinical objectives. Token-level supervision optimizes lexical overlap and local fluency rather than factual correctness and clinical reasoning [53]. It also induces bias, encourages hallucinations that read plausible but are unsupported by images, and offers weak guidance for factual coverage [68]. Consequently, CNN+LSTM models, Transformers, and even LVLMs can produce stylistically consistent but clinically incorrect reports, underscoring the need for learning signals that directly optimize clinical correctness instead of token-by-token imitation.

Building on this mismatch between token-level training and clinical goals, a growing line of work injects semantic supervision to better align text with image-grounded evidence. Contrastive learning aligns image-report pairs in a shared representa-

tion space; yet offering only global signals that overlook fine-grained phenomena and entity-relation structure [40, 95, 109]. Multitask learning jointly trains classification/localization with report generation, providing explicit semantic anchors that improve coverage and inhibit hallucinations. However, labels are incomplete or noisy, the categories are coarse, and extra heads push the model toward frequent findings and away from rare but critical ones [28, 92, 94]. Dynamic traceback learning partially advance the semantic consistency by masking or backtracking from generated tokens to visual evidence, improving robustness to spurious correlations; still, it relies on proxy curricula that do not directly reflect clinical quality [103]. To address these gaps, a semantic-driven reinforcement learning (SRL) method for medical report generation, termed MRG-R1, is proposed to directly optimize clinical correctness rather than token overlap. Specifically, a pretrained X-ray–specialized BERT model [80] is employed as the reward model to align generated reports with clinical annotations. Training is optimized using Group Relative Policy Optimization (GRPO) [78], a lightweight alternative to Proximal Policy Optimization (PPO) [76], which stabilizes learning via group-relative comparison and naturally encourages exploration of diverse description styles and latent Chain-of-Thought (CoT) without requiring explicit CoT annotations [14]. Concretely, the CheXbert-based margin-cosine reward (MCCS) delivers polarity-sensitive, sequence-level supervision that rewards per-finding agreement and overall coverage while sharply penalizing unsupported or contradictory statements, directly correcting the style–vs.–correctness mismatch of token-likelihood training. In parallel, GRPO's group-relative baseline amplifies the best candidate within each case without a learned value network, lowering variance and compute, and thus provides a stable, scalable training regime that remains clinically anchored.

The SRL framework is instantiated by transforming clinical semantics into a reward and optimizing the model using GRPO. For each study, a small group of candidate reports is generated from the current policy, and a clinical-efficacy reward is computed using CheXbert [80]. Specifically, CheXbert produces 14 disease labels for both the generated report and the reference, which are mapped to 14-dimensional vectors; a margin-based cosine similarity between the two vectors serves as the reward signal,

directly evaluating semantic agreement rather than n-gram overlap. To elicit explicit reasoning without annotated Chain-of-Thought (CoT) data, a brief system instruction is prepended, and a format reward is applied that grants additional points when the model adheres to a "Reasoning → Report" structure. The total reward combines the clinical-efficacy and format components in a weighted sum. Group Relative Policy Optimization (GRPO) [78] is then employed, normalizing rewards within each group using a groupwise baseline to form relative advantages, thereby producing low-variance updates that stabilize training compared with token-level objectives.

In this work, MRG-R1 is presented as a medical LVLM fine-tuned with GRPO to produce clinically aligned reports with explicit, self-generated reasoning. The main contributions are summarized as follows:

- **Semantic-Driven RL with GRPO for Clinically Aligned Medical Report Generation.** A semantic-driven reinforcement learning (SRL) framework is introduced for clinically aligned medical report generation, optimizing a clinically grounded reward and updating the policy via GRPO. This finetuning procedure enhances clinical alignment and elicits explicit reasoning without dependence on token-level supervision. The resulting model, **MRG-R1**, generates step-by-step rationales and radiology reports that adhere to clinical semantics.
- **CheXbert-Guided Clinical Efficacy Reward and Instruction-Driven Explicit Reasoning.** CheXbert's 14-label outputs for both generated and reference reports are converted into vectors, and a margin-based cosine-similarity reward is computed to provide a clinically grounded semantic supervision signal. Additionally, an instruction and format reward encourage a clear "Reasoning → Report" narrative without the need for annotated Chain-of-Thought (CoT) data.
- **Comprehensive Empirical Gains and Ablations on MIMIC-CXR and IU X-Ray.** Extensive experiments and ablation studies conducted on the IU X-Ray [17] and MIMIC-CXR [29] datasets validate the clinical efficacy and

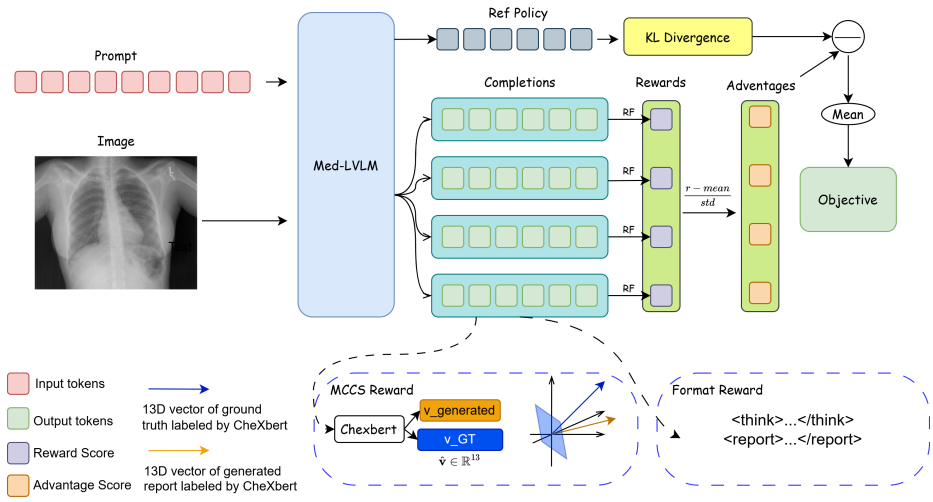


Figure 4.1 – Overview of SRL. For each study, the policy samples a group of candidate reports; a margin CheXbert cosine reward (MCCS) and a lightweight format reward are combined to compute group-relative advantages for GRPO updates under a KL constraint to a reference policy.

robustness of the proposed approach, demonstrating consistent improvements across standard evaluation metrics.

4.2 Method

Post-training for medical report generation is formulated as reinforcement learning over a clinically grounded reward, prioritizing report-level semantic fidelity rather than token likelihood. Concretely, the LVLM is fine-tuned using GRPO [78], a value-free, groupwise policy-gradient method chosen for its stability and computational efficiency when optimizing non-differentiable, report-level clinical rewards. The optimization signal is a CheXbert-guided reward that scores agreement across 14 chest-x-ray observations between the generated report and the reference, providing direct supervision on clinical content and reducing the reliance on n-gram overlap. The following sections detail (i) the GRPO training loop—covering sampling, group-relative advantage computation, and update rule—and (ii) the reward functions, including the CheXbert margin-cosine design and its aggregation at the report level.

4.2.1 Group Relative Policy Optimization (GRPO)

PPO [76] and GRPO [78] instantiate a family of post-training alignment algorithms that optimize reward-defined objectives rather than likelihood. This mechanism biases generation toward clinically aligned, report-level targets, providing direct supervision on semantic fidelity beyond token overlap. GRPO is closely related to PPO but differs in two key aspects: first, GRPO estimates the advantage using group-based estimation rather than a value function; second, it employs a set of fixed rules as the reward signal instead of a learned reward model.

Let $P(Q)$ denote the training set of inputs ("studies"); a single input is $q \in P(Q)$, $\pi_{\theta_{\text{old}}}$ and $\pi_{\theta_{\text{new}}}$ for the old policy (used to sample responses in the current update) and the current policy (parameters being optimized), respectively. A complete response o means the full generated report for q. A frozen reference policy, $\pi_{\theta_{\text{ref}}}$, is employed to regularize updates. Let G be the group size, the number of responses sampled per input q at each iteration, yielding $\{o_i\}_{i=1}^{G}$.

The GRPO objective is

$$
\begin{aligned}
J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} & \\
\Bigg[\frac{1}{G}\sum_{i=1}^{G} \min\Bigg(& \frac{\pi_{\theta_{\text{new}}}(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \\
& \text{clip}\Big(\frac{\pi_{\theta_{\text{new}}}(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)},\, 1-\epsilon,\, 1+\epsilon \Big) A_i \Bigg) \\
& - \beta\, D_{\text{KL}}\Big(\pi_{\theta_{\text{new}}} \,\Big\|\, \pi_{\theta_{\text{ref}}}\Big) \Bigg]
\end{aligned}
\tag{4.1}
$$

Here, the policy ratio $\frac{\pi_{\theta_{\text{new}}}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ measures how the new policy probability of o_i changes relative to the old policy; A_i is the estimated advantage for response o_i; $\epsilon > 0$ is the clipping threshold that limits overly large updates by replacing the raw ratio with its clipped version; and $D_{\text{KL}}(P\|Q)$ is the KL divergence [33] between the new policy and the reference policy, scaled by $\beta \geq 0$ to control policy drift. Intuitively, the "min" enforces the clipped surrogate familiar from PPO, while the KL term keeps the

updated policy close to $\pi_{\theta_{\text{ref}}}$.

Unlike PPO—which estimates A_i via a learned value function/critic—GRPO computes A_i within the sampled group for the same input to avoid value estimation. Concretely, with rewards $r_i = R(q, o_i)$ from our rule/labeler-based clinical reward R, a normalized, within-group advantage used

$$\begin{aligned} \bar{r} &= \frac{1}{G}\sum_{i=1}^{G} r_i, \\ \sigma_r &= \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \bar{r})^2 + \varepsilon}, \\ A_i &= \frac{r_i - \bar{r}}{\sigma_r}. \end{aligned} \tag{4.2}$$

where $\bar{r}$ and σ_r are the group mean and standard deviation, Here, $\varepsilon > 0$ is a small constant for numerical stability; when the group rewards are identical (variance near zero), set ε to a value on the order of 10^{-8}–10^{-6}. This relative construction compares candidates conditioned on the same study q. sharpening the learning signal for report-level clinical rewards without training a critic.

4.2.2 Reward Functions

Format Reward

A format reward is applied to elicit explicit, auditable reasoning without requiring CoT annotations. The prompt asks the model to place intermediate reasoning inside `<think>...</think>` and the final radiology report inside `<report>...</report>`. A rule-based scorer evaluates only structure: tags must be present, correctly ordered, well-formed (balanced), and non-empty. Outputs that fully comply receive a score of 1, with partial credit for minor violations; otherwise the score is 0. This term is added with a small weight relative to the clinical reward so optimization remains driven by medical correctness. Under GRPO's group-relative updates, candidates that satisfy the structure reliably obtain higher relative advantages within the same case, teaching

the policy to produce a stable two-stage "reasoning → report" format. The benefits are threefold: (i) decoupling thinking from the final narrative, (ii) improving readability and downstream parsing, and (iii) enabling auditability to localize hallucinations or inconsistencies.

Margin Chexbert Cosine Similarity Reward (MCCS)

Beyond enforcing output structure via the format reward, optimization is driven primarily by a clinically grounded signal that evaluates report-level semantics. This signal is instantiated as a Margin CheXbert Cosine Similarity (MCCS) reward, which converts CheXbert's 14-label [80] outputs into signed vectors and rewards their margin-calibrated cosine agreement, providing a continuous target for GRPO. For each study, CheXbert provides a 14-way multi-class label over common chest X-ray observations (e.g., Atelectasis ... No Finding). Each observation is then mapped to a scalar by

$$f(\text{pos}) = 1, f(\text{neg}) = -1, f(\text{uncertain}) = 1, f(\text{blank}) = 0. \tag{4.3}$$

and construct report-level vectors $\mathbf{z}(y), \mathbf{z}(y^\star) \in \mathbb{R}^{13}$ over the 13 disease-specific categories only (exclude No Finding) for the generated report y and the reference $y^\star$.

$$z_j(y) = f\Big(\ell_j(y)\Big), z_j(y^\star) = f\Big(\ell_j(y^\star)\Big), \quad j = 1, \ldots, 13. \tag{4.4}$$

Mapping uncertain to 1 treats hedged mentions as actionable suspicion rather than neutrality, which matches clinical practice: when radiologists hedge, they are flagging a possible abnormality that warrants attention. By contrast, blank is 0, reflecting true omission. This choice biases the reward toward sensitivity—it favors correctly surfacing potential findings and still penalizes polarity reversals (positive vs. negative) most strongly via the signed embedding. It also discourages "safe" under-calling: labeling everything as uncertain no longer evades penalties if the reference is negative (–1) or omitted (0), and it earns credit only when uncertainty aligns with a true or suspected abnormality. The No Finding dimension is also excluded from the cosine similarity.

In CheXbert [80], No Finding is typically set to 1 when all other disease labels are 0. As a result, it can dominate vector norms and inflate apparent agreement via complementarity, and it is highly sensitive to reporting style or templated omissions, thereby introducing noise. Removing this dimension focuses the signal on per-finding clinical agreement and avoids pseudo-alignment driven by a global catch-all label.

Then measure report-level agreement via cosine similarity

$$\mathrm{CCS}(y, y^\star) = \frac{\langle z(y), z(y^\star)\rangle}{\Big(\|z(y)\|_2 + \varepsilon\Big)\Big(\|z(y^\star)\|_2 + \varepsilon\Big)}, \quad \varepsilon = 10^{-8}. \tag{4.5}$$

This ε guarantees numerical safety even when one vector is (nearly) zero after preprocessing (e.g., with the No Finding dimension removed), while leaving values effectively unchanged when norms are in a normal range. To calibrate the signal and emphasize clinically meaningful improvements, The cosine similarity is converted into a margin-shaped reward:

$$\mathrm{MCCS}(y, y^\star, m) = \max\left(\frac{\mathrm{CCS}(y, y^\star) - m}{1 - m}, 0\right), m \in (-1, 1). \tag{4.6}$$

This piecewise-linear shaping has three advantages. (i) Margin filtering. Scores at or below m yield zero reward, suppressing weak alignments (e.g., incidental overlap) and focusing learning on clinically aligned matches. (ii) Dynamic-range normalization. The division by $(1 - m)$ maps $CCS \in [m, 1]$ to $[0, 1]$, ensuring comparable reward scales across studies and increasing within-group variance when m is moderate—beneficial for GRPO's group-relative advantages. (iii) Stable gradients. The linear slope $1/(1-m)$ avoids early saturation near high similarity and provides smooth, interpretable shaping; $MCCS = 1$ if and only if the two label vectors coincide up to positive scaling.

In all cases, MCCS acts as a continuous, clinically grounded reward at the report level, providing partial credit for near matches and stronger penalties for polarity mistakes than for uncertainty/omission, thereby aligning optimization with clinical correctness rather than token overlap.

4.3 Experimental Setup

4.3.1 Datasets

The experiments are conducted on the MIMIC-CXR [29] and IU X-Ray [17] datasets. MIMIC-CXR is the larger corpus, containing 473,057 chest X-ray images and 227,835 radiology reports, whereas IU X-Ray comprises 7,470 images and 3,955 reports. For fair comparison with prior work, the MIMIC-CXR split of approximately 222.8k/1.8k/3.3k samples for training, validation, and test is adopted following [10, 102]. For IU X-Ray, a 70/10/20 train/validation/test split is used following [10, 11]. Unless otherwise noted, multi-view studies (reports associated with multiple images) are treated as multiple image–report pairs, with each image paired to the same report and counted as a separate sample.

4.3.2 Implementation Details

All experiments are conducted on an AMD EPYC 7742 CPU with 64 cores and two NVIDIA A100 GPUs under CUDA 12.4. Finetuning is performed on HuatuoGPT-Vision-7B-Qwen2.5VL [1] [8], a Qwen2.5-VL [2] [2]-based vision–language model further aligned with medical image–text and instruction data. Training utilizes FlashAttention-2 [16] and bfloat16 mixed precision. For parameter-efficient tuning, LoRA [22] is applied with rank $r = 128$, scaling $\alpha = 256$, and dropout 0.05. Adapters are inserted into the attention query/key/value/output projections and the MLP gate/up/down projections. Optimization employs AdamW (8-bit) [52] with a learning rate of 5e-6, momentum terms $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay 0.1. A cosine learning-rate schedule with 10 DeepSpeed ZeRO-1 [70] is enabled for optimizer-state sharding and memory efficiency. For GRPO, each prompt yields four sampled completions, and group-relative advantages are computed against the reference policy. During training, decoding employs stochastic sampling with temperature = 1.0, top-p = 0.9, top-k

[1] https://huggingface.co/FreedomIntelligence/HuatuoGPT-Vision-7B-Qwen2.5VL
[2] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

= 50, and repetition penalty = 1.05. In the experiment combining multiple rewards, the total reward is defined as a weighted sum of the clinical and format rewards with coefficients 0.75 and 0.25, respectively. During RL post-training, we use a system prompt to specify the task and enforce a structured output format with `<think>` and `<report>` tags. Only the content inside `<report>` is used for reward computation and evaluation, while `<think>` is not evaluated and serves only as an optional scaffold. The full training prompt is listed in Appendix A (Listing A.1).

4.3.3 Evaluation Metrics

The quality of generated reports is evaluated using clinical efficacy (CE) metrics that emphasize factual correctness rather than stylistic similarity. Specifically, CheXbert-based precision, recall, and F1 scores are computed over 14 chest X-ray observations defined by CheXbert, following the standard evaluation protocol established in prior work [12, 80, 102, 103]. In contrast, conventional NLG metrics such as BLEU, ROUGE, and CIDEr primarily reward n-gram overlap and template reuse, which often obscure factual adequacy and fail to penalize polarity errors. Multiple studies have shown that such lexical metrics correlate weakly with radiologists' judgments of factual accuracy, whereas CE metrics better track clinically relevant errors [51, 61, 105]. At inference time, we adopt a minimal instruction prompt to request a radiology report, and the generated `<report>` text is used for metric computation. The exact evaluation prompt is provided in Appendix A (Listing A.2).

4.3.4 Baselines

In this study, MRG-R1 compared with three families of baselines in IU X-Ray and MIMIC-CXR, using code / checkpoints released when available and retraining with the authors' settings otherwise. (A) *Token-level MLE generators*: R2Gen [11] and R2GenCMN [9], representative encoder - decoder / transformer models trained under teacher forcing. (B) *Instruction-tuned medical LVLMs*: BioMedGPT [106], LLaVA-Med [37], CheXagent [12], HuatuoGPT-Vision [8], and MedGemma-4B/27B [77],

evaluated under a uniform prompting and decoding setup without additional fine-tuning on splits to probe zero / few shot reporting ability and domain alignment. (C) *Semantic supervision*: DTrace [103], DCL [39], GSKET [102], CXRMate [58], and RadFM [97], which inject clinical semantics through traceback, contrastive / matching, knowledge graphs or radiology-focused pre-training. Note that some baselines (e.g., CheXagent, MedGemma) are instruction-tuned on substantially broader medical corpora beyond MIMIC-CXR and IU X-Ray, and their performance may partly reflect pretraining coverage rather than architecture alone, therefore treat them as strong external baselines rather than strictly comparable models.

4.4 Results

4.4.1 Quantitative Analysis

In this study, the proposed method compared against various established report generation models across two datasets.

Across both datasets, MRG-R1 delivers strong clinical efficacy (CE) and is especially competitive on IU X-Ray. On IU X-Ray, MRG-R1 attains the highest F1=51.88, edging out classical encoder–decoder baselines such as R2GenCMN (50.53) and matching the top LVLM-style systems (e.g., CheXagent 51.15). The gains come from a balanced improvement in both precision (50.86) and recall (52.98), indicating that SRL with GRPO improves sensitivity to clinically salient findings while maintaining low false positives.

On MIMIC-CXR, MRG-R1 achieves F1=40.39, competitive with recent medical LVLMs (e.g., MedGemma-4B 41.08) and clearly above classic MLE baselines such as R2GenCMN (27.8). Notably, MRG-R1 exhibits higher precision (45.32) than MedGemma-4B (40.77) but somewhat lower recall (37.70 vs. 41.40), suggesting the clinically grounded reward reduces unsupported findings while remaining conservative on ambiguous cases. While CheXagent and MedGemma perform well, part of the gains may re-

Method	IU X-Ray			MIMIC-CXR		
	Precision	Recall	F1	Precision	Recall	F1
R2Gen [11]	50.60*	48.76*	46.99*	33.30	27.30	27.60
R2GenCMN [9]	50.00*	51.07*	50.53*	33.40	27.50	27.80
RadFM [97]	14.27*	11.93*	12.99*	10.03*	12.08*	10.96*
MedGemma-4B [77]	23.93*	22.83*	23.37*	40.77*	41.40*	**41.08***
MedGemma-27B [77]	15.40*	21.21*	17.84*	35.18*	36.95*	36.04*
BioMedGPT [106]	36.00*	35.40*	35.50*	29.00*	31.40*	28.60*
LLaVA-Med [37]	18.63*	23.37*	20.73*	26.99*	12.03*	16.64*
CheXagent [12]	50.37*	51.96*	51.15*	45.60*	24.59*	31.95*
HuatuoGPT-Vision [8]	5.87*	7.33*	6.52*	23.67*	16.51*	19.45*
DTrace [103]	–	–	–	41.10	**43.60**	39.10
DCL [39]	–	–	–	**47.10**	35.20	37.30
GSKET [102]	–	–	–	45.80	34.80	37.10
CXRMate [58]	28.30	35.10	27.70	43.80	34.90	35.70
MRG-R1	**50.86**	**52.98**	**51.88**	45.32	37.70	40.39

Table 4.1 – Clinical efficacy (CE) comparison on IU X-Ray and MIMIC-CXR. CE is computed with CheXbert over the 14 standard observations. * denote scores is not provided by author and reproduced by me. **Bold** marks the best performance per column; underline marks the second best. For IU X-Ray, CE is obtained by running CheXbert on both the generated and the ground-truth reports, consistent with prior work.

flect wider pretraining corpora beyond MIMIC-CXR/IU X-Ray rather than architecture alone; these models are therefore regarded as strong external baselines rather than strictly comparable counterparts. Compared with instruction-tuned or generalist LVLMs (e.g., LLaVA-Med, BioMedGPT), the CE advantage is substantial, underscoring the benefit of optimizing a clinical signal rather than relying solely on token-level imitation or generic instruction tuning.

It is also observed that methods injecting semantic signals without reinforcement learning (RL) exhibit limited improvements in clinical alignment (e.g., DTrace's traceback supervision; CXRMate's longitudinal semantic reward) yield stronger CE than early MLE systems, but MRG-R1 remains competitive or superior on average while using a lightweight, value-free GRPO objective. To further illustrate the effectiveness and stability of RL post-training, we report training dynamics including total reward

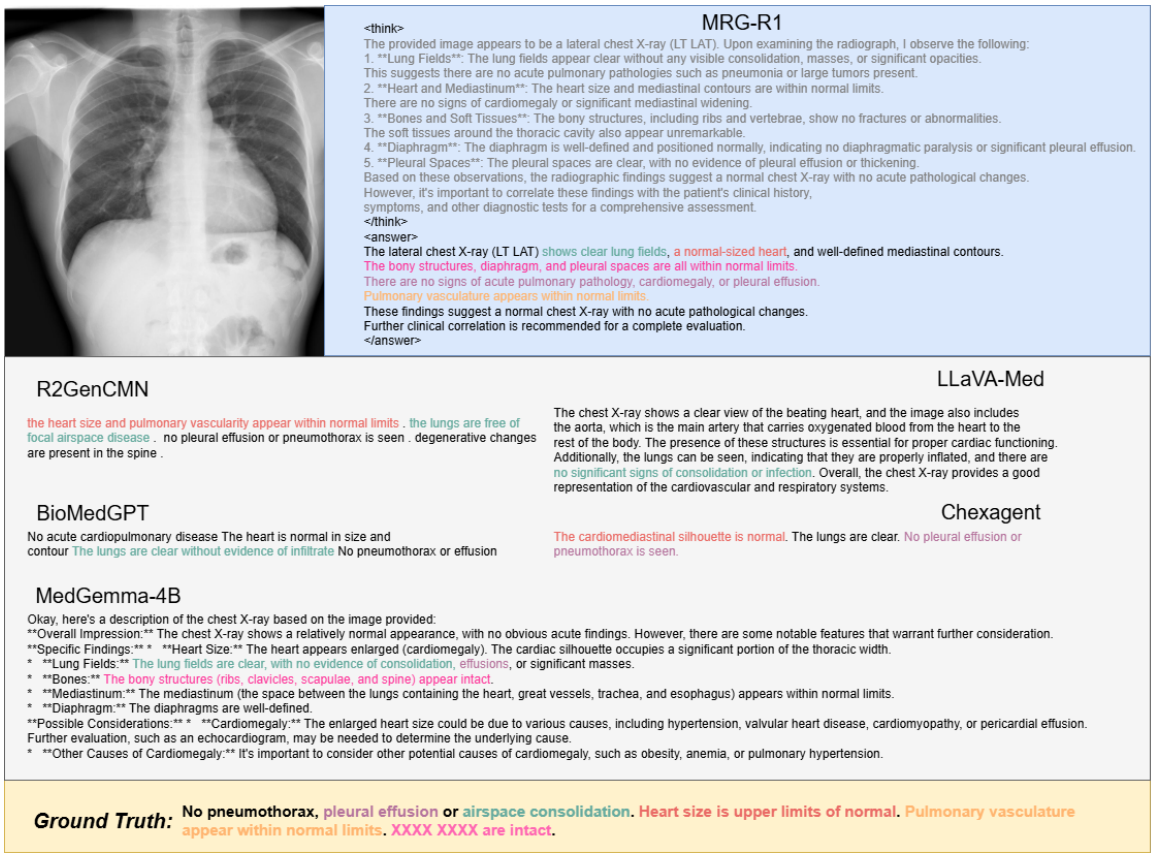


Figure 4.2 – IU X-Ray qualitative comparison (ID: $CXR2787_IM-1222$). Different text colors highlighting various medical descriptions for comparison with the Ground Truth.

(and its components), KL divergence to the reference policy, and completion length. The full training curves on IU X-ray and MIMIC-CXR are provided in Appendix A (Figures A.1 and A.2).

4.4.2 Qualitative Analysis

Beyond aggregate CE scores, report outputs from MRG-R1 are compared with strong baselines on representative IU X-Ray and MIMIC-CXR studies (Figures 4.2 - 4.3). The examples illustrate four recurrent dimensions that drive clinical utility: (i) polarity handling (affirmation/negation of key findings), (ii) uncertainty expression, (iii) omission vs. hallucination, and (iv) structural coherence.

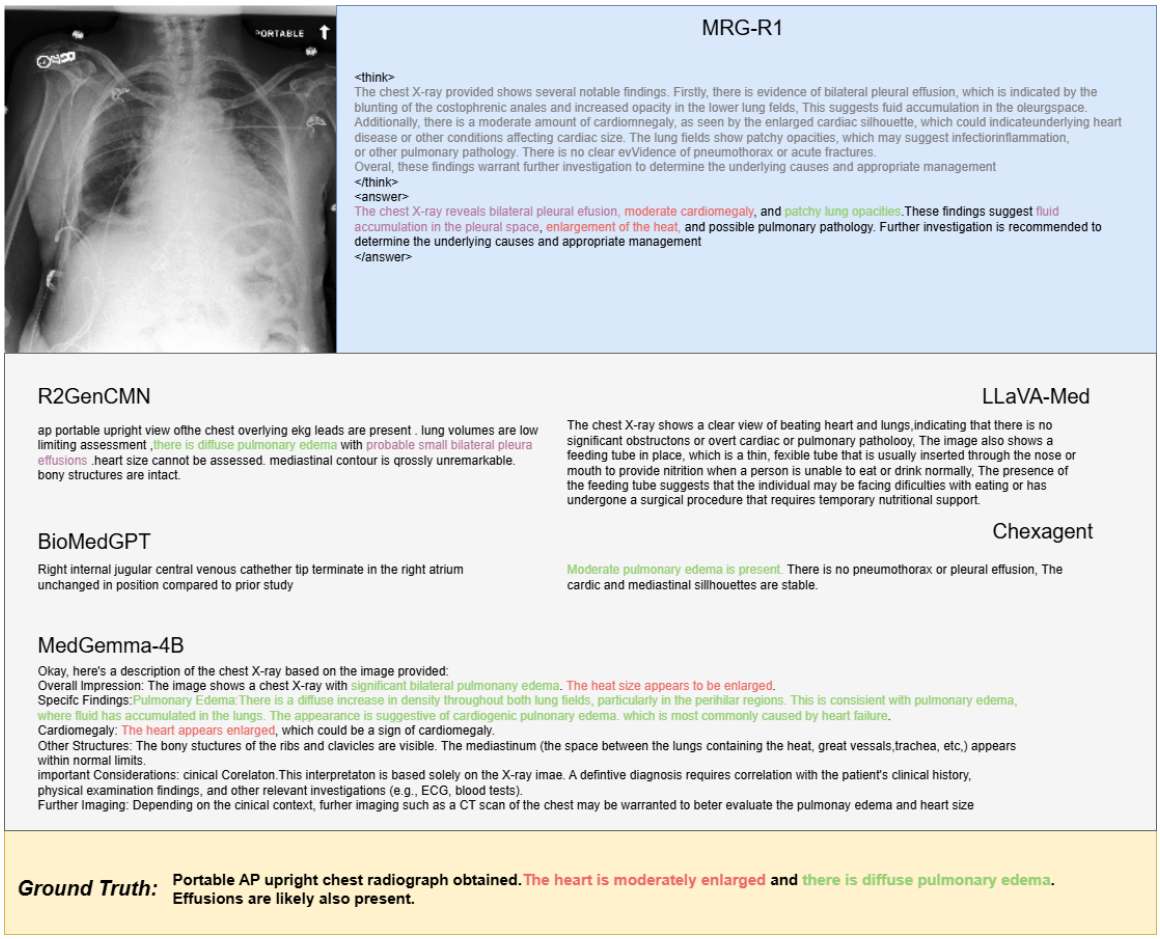


Figure 4.3 – MIMIC-CXR qualitative comparison (ID: $p16855430$). Different text colors highlighting various medical descriptions for comparison with the Ground Truth.

On IU X-Ray (Fig.4.2), the reference emphasizes normal lungs and pleura with a heart size at the upper limit of normal. MRG-R1's think→report format yields concise itemized statements that preserve correct negatives (no pneumothorax/effusion/-consolidation) and a near-normal cardiac size, aligning closely with the ground truth. Several baselines deviate: MedGemma-4B hallucinates cardiomegaly which is polarity error, R2GenCMN introduces extraneous skeletal comments which is hallucination, and instruction-tuned LVLMs (e.g., LLaVA-Med) generate fluent but generic prose that under-specifies required clinical elements.

On MIMIC-CXR (Fig.4.3), the report documents cardiomegaly, pulmonary edema, and likely effusions. MRG-R1 captures all three with consistent polarity, while CheXagent denies effusion which is polarity inversion, R2GenCMN hedges on heart size, and BioMedGPT emphasizes line positions while missing pathology (omission). MedGemma-4B identifies edema/cardiomegaly but is less reliable on effusions. These patterns mirror the quantitative findings: optimizing a polarity-sensitive, sequence-level reward (MCCS) plus a light reasoning format reduces unsupported positives/negatives, improves coverage of salient findings, and yields reports that are both clinically faithful and structurally auditable.

4.4.3 Ablation Studies

In this section, ablation studies are conducted on the proposed SRL using the IU X-Ray and MIMIC-CXR datasets to assess the contribution of each component within the method.

SRL is ablated on both datasets to analyze the contribution of each component (Table 4.2). (1) supervised finetuning (SFT, cross-entropy), (2) text-level NLG rewards (BLEU/ROUGE/CIDEr), (3) a format-only reward that enforces a `<think>` → `<report>` structure (Format), (4) a clinical reward via report-level CE-F1 (with-/without Format), and (5) the proposed margin CheXbert cosine similarity (MCCS, with/without Format). This sequence disentangles stylistic supervision, structural guidance, and clinically grounded objectives.

Method	IU X-Ray			MIMIC-CXR		
	Precision	Recall	F1	Precision	Recall	F1
Base	5.87	7.33	6.52	23.67	16.51	19.45
Base + SFT	3.86	7.09	4.99	24.27	15.00	14.64
Base + NLG	41.19	15.93	22.97	24.74	8.56	12.72
Base + CE F1	45.89	43.78	44.81	36.38	25.08	29.69
Base + reasoning	24.21	38.33	29.67	27.71	25.53	26.58
Base + CE F1 + reasoning	50.04	<u>52.73</u>	<u>51.35</u>	33.00	28.87	29.50
Base + MCCS	**53.27**	46.51	49.66	<u>36.07</u>	**44.69**	<u>38.67</u>
Base + MCCS + reasoning	<u>50.86</u>	**52.98**	**51.88**	**45.32**	<u>37.70</u>	**40.39**

Table 4.2 – Ablation on IU X-Ray and MIMIC-CXR starting from a zero-shot HuatuoGPT-Vision-7B (Base). Incrementally add: supervised finetuning (SFT, cross-entropy), text-level NLG rewards (BLEU+ROUGE+CIDEr), a format-only reward enforcing a `<think>` → `<report>` structure (Format), a clinical reward via report-level CE-F1 (with/without Format), and the margin CheXbert cosine similarity (MCCS, with/without Format). **Bold** marks the best per column; <u>underline</u> the second best.

Relative to Base, optimizing purely lexical NLG rewards (+NLG) improves fluency but yields limited clinical efficacy (CE): F1 rises only to 22.97 on IU X-Ray and 12.72 on MIMIC-CXR, consistent with the weak linkage between n-gram overlap and factual correctness. Replacing the objective with a clinical signal (+CE-F1) substantially improves CE (IU 44.81; MIMIC 29.69), indicating that label-consistency supervision reduces polarity errors and under-calling. A format-only constraint (+Format) increases recall (IU 38.33; MIMIC 25.53) at some cost to precision, while +CE-F1+Format stabilizes negation/uncertainty templates and recovers a strong precision–recall balance (IU F1 51.35).

MCCS is the most effective shaping in this setting. Compared with CE-F1, MCCS maps CheXbert labels to signed vectors (pos = 1, neg = -1, blank = 0, uncertain = 1), excludes the catch-all *No Finding*, and applies a margin that suppresses weak matches. This polarity-sensitive, sequence-level signal widens case-level score separation, which GRPO's group-relative updates leverage to amplify the best candidate per study. Empirically, +MCCS boosts recall on MIMIC-CXR (44.69; second-best F1 38.67), and +MCCS+Format delivers the best overall CE on both datasets (IU F1 51.88; MIMIC F1 40.39) with the highest MIMIC precision (45.32). These trends support MCCS as a stronger clinical reward than CE-F1 under GRPO and motivate pairing it with a light format constraint for stable long-form generation.

4.5 Discussion

4.5.1 Clinical Significance of Quantitative Gains

Table 4.1 shows that optimizing a clinically grounded signal yields consistent CE improvements across datasets, achieving a strong precision–recall balance on IU X-Ray and competitive results on MIMIC-CXR. These findings confirm that integrating GRPO, MCCS, and reasoning supervision enhances clinical usefulness.

Such gains arise from a reward design that explicitly encodes clinical semantics and polarity awareness, enabling the model to reason about findings rather than mimic lexical patterns. The MCCS reward defines a polarity-sensitive, sequence-level objective: CheXbert labels are mapped to signed vectors (pos=1, neg=–1, blank=0, uncertain=1), which penalizes polarity errors most strongly while treating clinically appropriate uncertainty as actionable. Excluding "No Finding" prevents a catch-all dimension from inflating apparent agreement, and the margin shape suppresses weak overlaps while expanding the useful dynamic range. Under GRPO's group-relative updates, this larger within-case variance helps the policy prefer the best candidate per study and supports stable updates, so the reward targets case-level coverage and polarity rather than token overlap.

A lightweight reasoning format (`<think>...</think>` → `<report>...</report>`) further encourages self-generated intermediate reasoning without human-annotated CoT. In GRPO's groupwise comparisons, well-formed outputs tend to receive higher rewards, helping the policy consistently produce coherent, auditable finding lists with stabilized negation/uncertainty templates.

Overall, these findings demonstrate that semantic supervision delivers stronger clinical alignment than MLE or SFT objectives. Whereas token-level training distributes local credit based on lexical similarity, semantic rewards provide global, polarity-aware feedback that suppresses hallucinations and improves finding coverage. This behavior aligns with prior evidence that lexical metrics correlate weakly with factual accuracy in radiology reporting.

4.5.2 Qualitative Insights and Error Taxonomy

Building on the qualitative examples above, model behaviors are synthesized into a broader taxonomy of strengths and error modes. MRG-R1 consistently demonstrates four qualitative improvements across the case studies in Figures 4.2–4.3: robust polarity control that reduces internal contradictions; broader coverage of salient findings beyond generic normal templates; calibrated use of uncertainty that avoids over-assertion; and clearer, auditable structure induced by the from think to report format. Together, these behaviors reflect a shift from pattern-based text imitation toward reasoning-driven clinical articulation.

Despite these advances, several residual errors persist. The most frequent are omissions of subtle or localized abnormalities, occasional polarity inversions in challenging studies, incomplete handling of devices or lines not covered by the reward schema, and conservative summaries on ambiguous cases. These limitations reflect how the current reward design and inputs shape model behavior.

These patterns trace to a clinically grounded but coarse reward space (14 labels), single-timepoint and variable-quality inputs, and the group-relative update that cannot recover cues missed by all sampled candidates. Mitigations include expanding re-

wards to device and localized labels, adding multi-view and prior-study conditioning, lightly calibrating confidence (penalizing over-certainty, rewarding warranted hedges), applying simple decoding constraints tied to auxiliary detectors, and targeted augmentation plus radiologist-in-the-loop preference updates for borderline findings.

Overall, sequence-level, polarity-aware optimization with light structural guidance yields clinically readable reports with fewer contradictions; addressing the coarse reward coverage and ambiguous-evidence cases should further translate these gains into dependable clinical utility.

4.5.3 Ablation-Driven Design Guidance

Table 4.2 summarizes ablations starting from a zero-shot HuatuoGPT-Vision-7B and progressively adding: (1) supervised finetuning (SFT, cross-entropy), (2) text-level NLG rewards (BLEU/ROUGE/CIDEr), (3) a format-only reward enforcing a `<think>` → `<report>` structure, (4) clinical rewards via report-level CE-F1 (with-/without format), and (5) margin CheXbert cosine similarity (MCCS) reward (with-/without format). This sequence isolates stylistic supervision, structural guidance, and clinically grounded objectives.

First, the format constraint is complementary. Used alone, it tends to increase recall at a modest cost to precision—encouraging the model to surface more candidate findings. Combined with a clinical reward (+CE-F1+format), it stabilizes negation/uncertainty templates and yields a better precision–recall balance, especially on IU X-Ray.

Second, MCCS is the most effective shaping in default setting. Compared with CE-F1, MCCS maps CheXbert labels to signed vectors (pos=1, neg=–1, blank=0, uncertain=1) and applies a margin that suppresses weak matches while excluding the catch-all "No Finding"dimension. This polarity-sensitive, sequence-level signal produces larger case-level score separation, which GRPO leverages in its group-relative updates to amplify the best candidate per study. Empirically, this yields stronger

recall on MIMIC-CXR (+MCCS) and the best overall F1 when coupled with the format constraint (+MCCS+format).

Taken together, these trends suggest a practical recipe: prioritize a clinically grounded reward; add lightweight format guidance to stabilize long-form outputs; and favor margin-shaped, polarity-aware objectives (MCCS) when using GRPO, as they create clearer within-case preference signals that translate into consistent CE gains (Table 4.2).

4.6 Conclusion

This study addresses the gap between token-level objectives and clinical goals in medical report generation by introducing a semantic-driven reinforcement learning (SRL) method that fine-tunes LVLMs with Group Relative Policy Optimization (GRPO) and a Margin CheXbert Cosine Similarity (MCCS) reward, complemented by a lightweight format reward. On MIMIC-CXR and IU X-Ray, the resulting MRG-R1 system achieves consistent gains on CheXbert-based precision, recall, and F1, with fewer polarity mistakes and fewer unsupported statements; training remains stable and compute-efficient due to GRPO's group-relative advantages and value-free updates. Ablations validate the contribution of medical-domain initialization, MCCS shaping, and the format constraint.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

This thesis addressed two persistent barriers to clinically useful medical report generation: (i) incomplete coverage of salient findings, often skewed toward "normal" and (ii) weak alignment between token-level training objectives and clinically grounded correctness. To overcome these gaps, two complementary contributions were presented. MRGAgents reframes reporting as a disease-specialized, multi-agent process in which each agent generates a concise, category-focused sentence, yielding more systematic coverage and mitigating normality bias. MRG-R1 recasts post-training as reinforcement learning over a clinically grounded, report-level reward optimized with GRPO, reducing polarity errors and unsupported statements while maintaining fluency. Evaluations on standard chest X-ray benchmarks demonstrated consistent gains in clinical efficacy alongside competitive language quality, indicating tangible benefits for triage and drafting.

Beyond raw scores, the two contributions offer a coherent recipe for clinically aligned generation. MRGAgents improves what is covered—breadth and recall across disease categories—by enforcing a checklist-like structure that maps naturally to radiology reading patterns. MRG-R1 improves how it is stated—factuality, calibrated uncertainty, and fewer hallucinations by optimizing a polarity-aware, sequence-level signal

rather than n-gram overlap. Together, they enhance auditability: sentence-scoped outputs are easier to verify, and a "reasoning → report" discipline clarifies evidential grounding.

This work remains scoped to chest X-ray reporting and leverages automatic labelers for supervision and evaluation, which can introduce noise and leave device/localization details under-specified; multi-agent aggregation, while parallelizable, also benefits from stronger coordination to preserve cross-sentence coherence. Within these bounds, the thesis advances a principled stance for safety-critical generation: design for comprehensive coverage, then optimize for clinically grounded correctness. The resulting synthesis, task decomposition for completeness plus clinically aligned reinforcement learning for correctness offers a practical path toward more reliable, workflow-compatible reporting systems.

5.2 Future Work

From a methodological perspective, the objective is to progress from "style imitation" to verifiable clinical reasoning. This entails shaping training objectives with structured rewards that explicitly encode entities, relations, anatomic locations, and polarities, aligning optimization with the factual schema radiologists rely on. In parallel, preference learning: from expert pairwise comparisons and multi-criteria judgments can complement rule-based rewards to better capture trade-offs among coverage, correctness, and clarity. Within the multi-agent paradigm, the framework evolves from parallel specialists to cooperative agents that exchange evidence, resolve conflicts, and reach consensus under a lightweight global controller. Finally, knowledge distillation will be explored to compress multi-agent behavior into a single deployable model, while uncertainty modeling and calibration (including selective generation, risk-sensitive decoding, and confidence-aware scoring) will be integrated so that the system can "acknowledge uncertainty" and propagate calibrated confidence throughout training and evaluation loops.

On the task and data side, the scope will be broadened beyond single-study chest ra-

diographs to encompass multi-modal, multi-temporal, and multi-organ settings—spanning CT, MRI, and ultrasound, as well as multi-view X-rays and longitudinal follow-up. Textually, the scope extends from Findings to Impression, Comparison, and Recommendations, thereby strengthening the linkage between evidence and clinical decision support. This will be paired with richer supervision (e.g., region-level grounding, lesion attributes, temporal change labels) and evidence-linked generation that highlights where each statement is supported in the image(s), improving traceability and auditability.

In summary, future work will (i) re-center optimization on structured, preference-informed, and uncertainty-aware clinical reasoning; (ii) advance multi-agent collaboration and distill it into compact models; and (iii) expand modalities, temporal context, and clinical sections to close the gap between fluent reporting and clinically verified, evidence-grounded reporting. These directions are mutually reinforcing and chart a path from accurate sentences toward trustworthy, end-to-end radiology assistance.

List of References

[1] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. (2023). Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

[2] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. (2025). Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

[3] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

[4] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

[5] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M. P., Nori, A., Alvarez-Valle, J., and Oktay, O. (2023). Learning to exploit temporal structure for biomedical vision–language processing.

[6] Booth, T. C. and Boele, F. W. (2025). Reporting turn-around-times are growing: interval scanning anxiety and other harms need mitigation. *European Radiology*, pages 1–2.

[7] Bustos, A., Pertusa, A., Salinas, J.-M., and De La Iglesia-Vaya, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

[8] Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G. H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al. (2024a). Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

[9] Chen, Z., Shen, Y., Song, Y., and Wan, X. (2021). Generating radiology reports via memory-driven transformer. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

[10] Chen, Z., Shen, Y., Song, Y., and Wan, X. (2022). Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*.

[11] Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

[12] Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J. M. J., Youssef, A., Cohen, J. P., Reis, E. P., et al. (2024b). Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

[13] Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., and Bansal, M. (2022). Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

[14] Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q. V., Levine, S., and Ma, Y. (2025). Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

[15] Dai, D., Zhang, Y., Xu, L., Yang, Q., Shen, X., Xia, S., and Wang, G. (2024). Pa-llava: A large language-vision assistant for human pathology image understanding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3138–3143. IEEE.

[16] Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

[17] Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

[18] Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. (2024). Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

[19] Guo, D. and Terzopoulos, D. (2024). Prompting medical large vision-language models to diagnose pathologies by visual question answering. *arXiv preprint arXiv:2407.21368*.

[20] Hong, J., Lee, N., and Thorne, J. (2024a). Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

[21] Hong, S., Xiao, L., Zhang, X., and Chen, J. (2024b). Argmed-agents: Explainable clinical decision reasoning with llm disscusion via argumentation schemes. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5486–5493. IEEE.

[22] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

[23] Huang, S.-C., Shen, L., Lungren, M. P., and Yeung, S. (2021). Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3942–3951.

[24] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

[25] Izhar, A., Idris, N., and Japar, N. (2025). Medical radiology report generation: A systematic review of current deep learning methods, trends, and future directions. *Artificial intelligence in medicine*, page 103220.

[26] Jain, S., Agrawal, A., Saporta, A., Truong, S. Q., Duong, D. N., Bui, T., Chambon, P., Zhang, Y., Lungren, M. P., Ng, A. Y., et al. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

[27] Jeong, J., Tian, K., Li, A., Hartung, S., Adithan, S., Behzadi, F., Calle, J., Osayande, D., Pohlen, M., and Rajpurkar, P. (2024). Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, pages 978–990. PMLR.

[28] Jing, B., Xie, P., and Xing, E. (2017). On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

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

[30] Kasalak, Ö., Alnahwi, H., Toxopeus, R., Pennings, J. P., Yakar, D., and Kwee, T. C. (2023). Work overload and diagnostic errors in radiology. *European Journal of Radiology*, 167:111032.

[31] Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., and Park, H. W. (2024). Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[32] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

[33] Kullback, S. (1951). Kullback-leibler divergence. *Tech. Rep.*

[34] Kwee, T. C. and Kwee, R. M. (2021). Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence. *Insights into imaging*, 12(1):88.

[35] Lee, S., Youn, J., Kim, H., Kim, M., and Yoon, S. H. (2025). Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, pages 1–13.

[36] Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z., et al. (2024a). Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*.

[37] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023a). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

[38] Li, C. Y., Liang, X., Hu, Z., and Xing, E. P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6666–6673.

[39] Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., and Chang, X. (2023b). Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

[40] Li, M., Lin, H., Qiu, L., Liang, X., Chen, L., Elsaddik, A., and Chang, X. (2024b). Contrastive learning with counterfactual explanations for radiology report generation. In *European Conference on Computer Vision*, pages 162–180. Springer.

[41] Li, Y., Liang, X., Hu, Z., and Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31.

[42] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

[43] Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., and Xie, W. (2023). Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

[44] Liu, F., Wu, X., Ge, S., Fan, W., and Zou, Y. (2021). Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.

[45] Liu, G., Hsu, T., McDermott, M., Boag, W., Weng, W., Szolovits, P., and Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. corr. *arXiv preprint arXiv:1904.02633*.

[46] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023a). Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

[47] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023b). Visual instruction tuning.

[48] Liu, J., Wang, Z., Ye, Q., Chong, D., Zhou, P., and Hua, Y. (2023c). Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

[49] Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881.

[50] Liu, X., Xin, J., Shen, Q., Huang, Z., and Wang, Z. (2025). Automatic medical report generation based on deep learning: A state of the art survey. *Computerized Medical Imaging and Graphics*, page 102486.

[51] Liu, Y., Wang, Z., Li, Y., Liang, X., Liu, L., Wang, L., and Zhou, L. (2024). Mrscore: Evaluating radiology report generation with llm-based reward system. *arXiv preprint arXiv:2404.17778*.

[52] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[53] Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

[54] Miura, Y., Zhang, Y., Tsai, E. B., Langlotz, C. P., and Jurafsky, D. (2020). Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*.

[55] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. (2023). Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

[56] Moratelli, N., Caffagni, D., Cornia, M., Baraldi, L., and Cucchiara, R. (2024). Revisiting image captioning training paradigm via direct clip-based optimization. *arXiv preprint arXiv:2408.14547*.

[57] Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T., Dinh, D. H., et al. (2022). Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429.

[58] Nicolson, A., Dowling, J., Anderson, D., and Koopman, B. (2024). Longitudinal data and a semantic similarity reward for chest x-ray report generation. *Informatics in Medicine Unlocked*, 50:101585.

[59] Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., et al. (2025). Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*.

[60] Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

[61] Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A. E., Moseley, M., Langlotz, C., Chaudhari, A. S., et al. (2024). Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*.

[62] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

[63] Pang, T., Li, P., and Zhao, L. (2023). A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, 22(1):48.

[64] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[65] Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., and Lu, Z. (2018). Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

[66] Qin, H. and Song, Y. (2022). Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458.

[67] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

[68] Ramesh, V., Chi, N. A., and Rajpurkar, P. (2022). Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR.

[69] Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2015). Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

[70] Rasley, J., Rajbhandari, S., Ruwase, O., and He, Y. (2020). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

[71] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

[72] Ritchie, B., Summerville, L., Sheng, M., Choi, M., Tirumani, S., and Ramaiya, N. (2025). Impact of turnaround time in radiology: The good, the bad, and the ugly. *Current Problems in Diagnostic Radiology*.

[73] Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. (2015). Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447.

[74] Sarto, S., Barraco, M., Cornia, M., Baraldi, L., and Cucchiara, R. (2023). Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[75] Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., and Moor, M. (2024). Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

[76] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[77] Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al. (2025). Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

[78] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

[79] Shi, W., Xu, R., Zhuang, Y., Yu, Y., Zhang, J., Wu, H., Zhu, Y., Ho, J., Yang, C., and Wang, M. D. (2024). Ehragent: Code empowers large language models for complex tabular reasoning on electronic health records. *arXiv preprint arXiv:2401.07128*.

[80] Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., and Lungren, M. P. (2020). Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

[81] Soleimani, M., Seyyedi, N., Ayyoubzadeh, S. M., Kalhori, S. R. N., and Keshavarz, H. (2024). Practical evaluation of chatgpt performance for radiology report generation. *Academic Radiology*, 31(12):4823–4832.

[82] Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., Zhang, Y., Li, H., and Yang, L. (2024). Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5034–5042.

[83] Takada, T., Suzuki, Y., Takushima, H., Tanoue, H., Sato, H., Kumar, A., Nishihara, H., Hori, T., and Ueki, K. (2024). Direct metric optimization for image captioning through reward-weighted augmented data utilization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8333–8346.

[84] Tang, X., Shao, D., Sohn, J., Chen, J., Zhang, J., Xiang, J., Wu, F., Zhao, Y., Wu, C., Shi, W., et al. (2025). Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

[85] Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. (2023). Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

[86] Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

[87] Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. (2024). Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

[88] Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

[89] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[90] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. (2024a). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

[91] Wang, X., Figueredo, G., Li, R., Zhang, W. E., Chen, W., and Chen, X. (2024b). A survey of deep learning-based radiology report generation using multimodal data. *arXiv preprint arXiv:2405.12833*.

[92] Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058.

[93] Wang, X., Wang, F., Li, Y., Ma, Q., Wang, S., Jiang, B., and Tang, J. (2025). Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5123–5133.

[94] Wang, Z., Han, H., Wang, L., Li, X., and Zhou, L. (2022). Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging*, 41(10):2803–2813.

[95] Wang, Z., Zhou, L., Wang, L., and Li, X. (2021). A self-boosting framework for automated radiographic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2433–2442.

[96] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

[97] Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., and Xie, W. (2025). Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

[98] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

[99] Xu, W., Chan, H. P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al. (2025). Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

[100] Yan, S., Cheung, W. K., Chiu, K., Tong, T. M., Cheung, K. C., and See, S. (2023). Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging*, 42(8):2211–2222.

[101] Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., and Xiao, L. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

[102] Yang, S., Wu, X., Ge, S., Zhou, S. K., and Xiao, L. (2022). Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

[103] Ye, S., Meng, M., Li, M., Feng, D., Naseem, U., and Kim, J. (2024). Dynamic traceback learning for medical report generation. *arXiv preprint arXiv:2401.13267*.

[104] Young, A., Wang, K. E., Jin, M. X., Avilla, K., Gilotra, K., Nguyen, P., and Ros, P. R. (2025). A hands-free approach with voice to text and generative artificial intelligence: Streamlining radiology reporting. *Journal of the American College of Radiology*, 22(2):200–203.

[105] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., Fonseca, E. K. U. N., Lee, H. M. H., Abad, Z. S. H., Ng, A. Y., et al. (2023). Evaluating progress in automatic chest x-ray radiology report generation. patterns 4, 9 (2023).

[106] Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B. D., Ren, H., et al. (2024). A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

[107] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

[108] Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. (2022). Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR.

[109] Zhao, R., Wang, X., Dai, H., Gao, P., and Li, P. (2023). Medical report generation based on segment-enhanced contrastive representation learning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 838–849. Springer.

[110] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Appendix A

Appendix

A.1 Implementation Details

A.1.1 Prompt Design of MRGAgents

The MRGAgents does not introduce a new or disease-specific prompt design. Instead, all agents follow the default prompt setting in the original BioMedGPT captioning/report-generation pipeline (e.g., the fixed captioning instruction used in BioMedGPT data construction, such as "what does the image describe?" in the GPT2BPE setting). The agent specialization in MRGAgents is achieved through data partitioning rather than prompt modification.

A.1.2 Prompt Design in Training Process of MRG-R1

For RL post-training, we use a system prompt to define the task and to enforce a structured output format. The model is instructed to produce an intermediate reasoning block and a final answer using `<think>` and `<report>` tags. We note that only the `<report>` content is used for reward computation and evaluation, while `<think>` is not evaluated and serves only as an optional scaffold during generation.

Listing A.1 – System prompt used for MRG-R1 training

```
Below is an instruction that describes a task, paired with an input that
    provides further context. Write a response that appropriately
    completes the request.Before answering, think carefully about the
    question and create a step-by-step chain of thoughts to ensure a
    logical and accurate response.

### Instruction:
You are a medical expert with advanced knowledge in clinical reasoning,
    diagnostics, and medical report generation.Please answer the following
    medical question based on the input image. Output the thinking process
    in <think> </think> and final answer in <report> </report> tags.The
    output answer format should be as follows:
<think> ... </think>
<report> ... </report>
```

A.1.3 Prompt Design in Evaluation Process of MRG-R1

During evaluation/inference, we adopt a minimal instruction prompt to request a radiology report, and the generated report text is used for computing clinical efficacy metrics. No additional constraints beyond the standard decoding settings are imposed unless otherwise stated.

Listing A.2 – System prompt used for MRG-R1 evaluation

```
Please generate a radiology report.
```

A.2 Additional Results

A.2.1 Training Curve of MRG-R1

Figures A.1 and A.2 report the training dynamics of MRG-R1 during RL post-training on IU X-ray and MIMIC-CXR, respectively. The total reward is computed as a weighted combination of the MCCS clinical reward and a lightweight format reward, while the KL divergence to the reference policy is monitored to ensure controlled updates. The x-axis denotes optimization steps (one policy update per step), and completion length is measured in generated tokens. Overall, the curves show rapid early improvements in reward followed by stabilization, with an early saturation of the format reward and a plateauing KL divergence, indicating stable convergence under KL regularization.

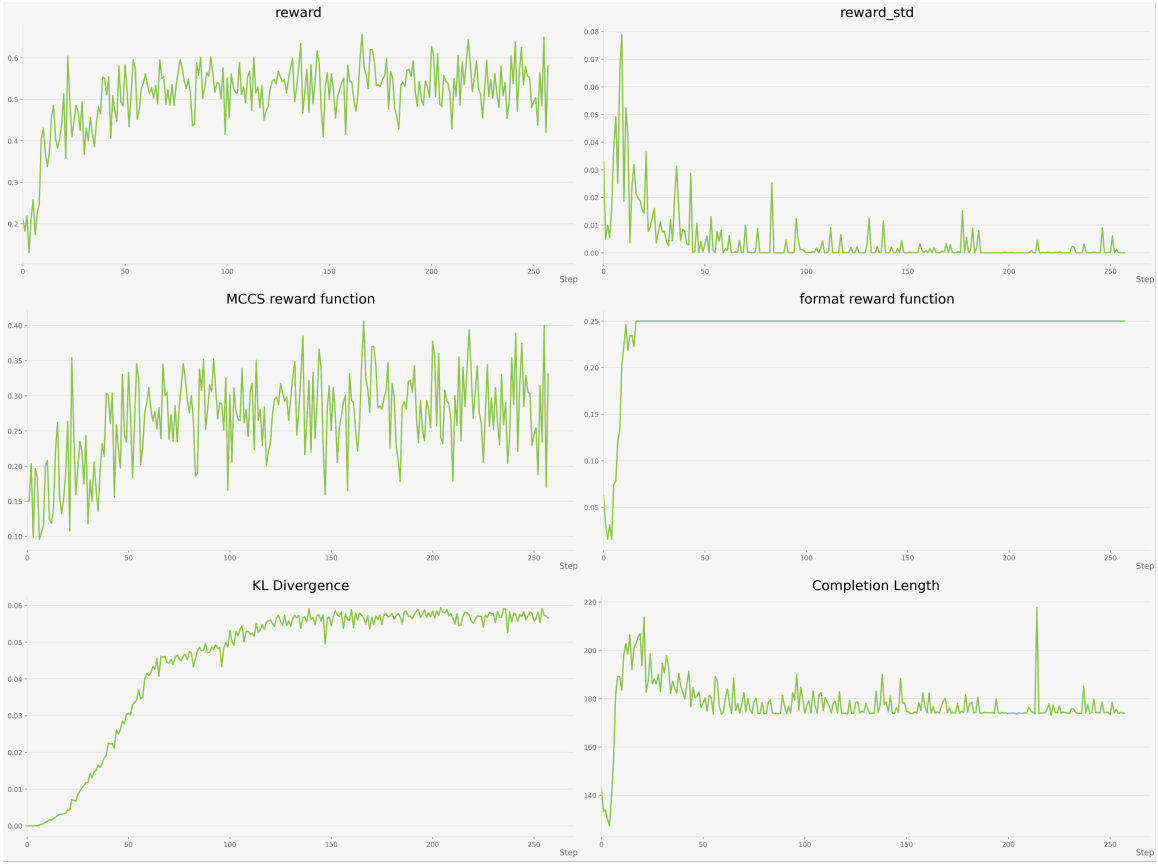


Figure A.1 – Training curves of MRG-R1 on IU X-ray. It shows the total reward (and reward standard deviation), the MCCS reward and the format reward used in the composite objective, the KL divergence between the current policy and the reference model, and the average completion length (in tokens) over training steps, illustrating stable optimization and convergence under KL regularization.

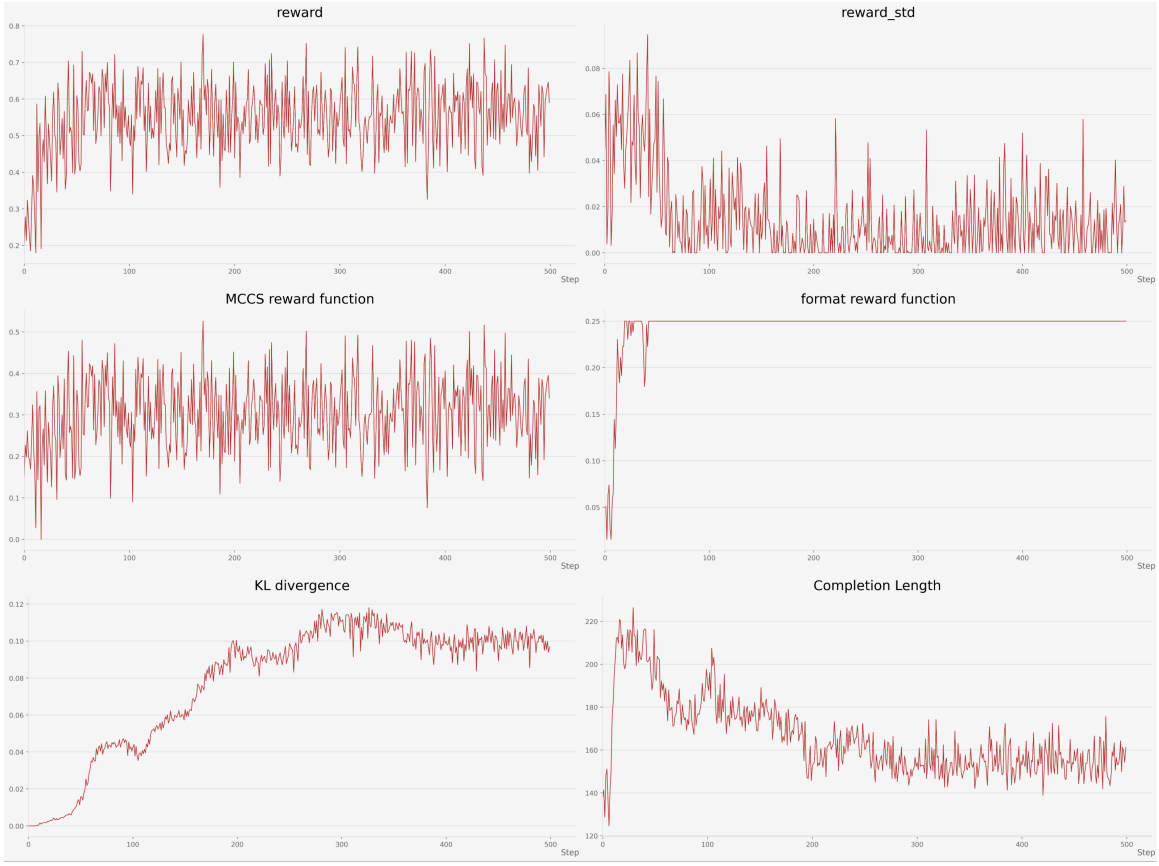


Figure A.2 – Training curves of MRG-R1 on MIMIC-CXR. It shows the total reward (and reward standard deviation), the MCCS reward and the format reward used in the composite objective, the KL divergence between the current policy and the reference model, and the average completion length (in tokens) over training steps, illustrating stable optimization and convergence under KL regularization.