MULTIMODAL CHEST DISEASE DIAGNOSIS SYSTEM

Krishna Singh Rajput, Shivani Chauhan, Nishtha Wagh, Kanishk Tanotra, Sharavanee Dommaraju Padmanabhan

ABSTRACT

Accurate interpretation of chest radiographs is critical for diagnosing thoracic conditions but remains a complex, expertise-driven task, especially in high-volume or resource-limited settings. This project presents the Multimodal Chest Disease Diagnosis System, an AI-driven framework that generates structured, radiologist-style reports directly from chest X-ray images.

Our approach leverages a vision-language model fine-tuned on a hybrid dataset composed of expert-authored and synthetically generated reports. To enhance factual consistency, the model was further optimized using reinforcement learning techniques that reward alignment between image content and textual output.

Initial evaluations using natural language generation metrics indicate strong performance in generating coherent and diagnostically relevant narratives, demonstrating the potential of this system as a foundation for AI-assisted radiology tools and educational platforms.

Keywords: Chest X-ray, Radiology Report Generation, Vision-Language Model, Artificial Intelligence, Reinforcement Learning, Structured Reporting, Synthetic Data Augmentation

1. INTRODUCTION

Chest radiography (CXR) is one of the most widely used diagnostic imaging techniques, playing a key role in detecting conditions such as pneumonia, pleural effusion, cardiomegaly, and pulmonary edema. Despite its clinical significance, interpreting chest X-rays remains a challenging task that demands expert knowledge, particularly in cases involving subtle findings or overlapping abnormalities.

These challenges are further amplified in high-volume clinical settings and resource-limited environments, where access to trained radiologists may be constrained. Additionally, radiology trainees often lack immediate feedback during interpretation, making the learning process both difficult and error-prone.

Traditional AI solutions, particularly Convolutional Neural Networks (CNNs), have made progress in automating CXR analysis through multi-label classification. However, these models typically produce fixed disease tags, lacking the contextual depth and narrative reasoning that radiologists provide in real-world reports. Their outputs are difficult to interpret and integrate into clinical workflows, limiting trust and applicability.

To address these limitations, recent research has turned toward Vision-Language Models (VLMs), which combine visual understanding with natural language generation. These models have shown promise in generating free-form, human-like reports from medical images—offering greater transparency, interpretability, and potential value in educational and diagnostic settings.

In this project, we present the Multimodal Chest Disease Diagnosis System, an AI-driven framework that generates structured radiology reports directly from chest X-ray images. By fine-tuning a large-scale vision-language model on a hybrid dataset consisting of real and synthetic radiology reports, we aim to produce outputs that reflect radiologist-style narratives while maintaining factual consistency. The system is further optimized using reinforcement learning techniques to ensure alignment between image content and generated text.

This report details the motivation, design, and evaluation of our system, highlighting its potential as a foundation for scalable, explainable, and educationally useful AI tools in radiology.

2. RELATED WORK

Automated interpretation of chest X-rays has been an active area of research in medical AI. Prior work in this domain generally falls into three categories: (i) convolutional neural networks (CNNs) for disease classification, (ii) vision-language models for joint image-text learning, and (iii) automated radiology report generation. Each has contributed to the field's advancement but also presents limitations that this project aims to address.

2.1 CNN-Based Disease Classification

CNNs have been widely adopted for chest X-ray diagnosis through multi-label classification. CheXNet [3], for example, used a DenseNet-121 architecture trained on ChestX-ray14 to detect pneumonia at a level comparable to radiologists. CheXpert [3] later introduced uncertainty-aware labeling for 14 thoracic conditions, helping models manage ambiguous cases. While CNNs perform well in predicting pathologies, they offer limited contextual information and lack explainability—failing to emulate the narrative reasoning used by human radiologists.

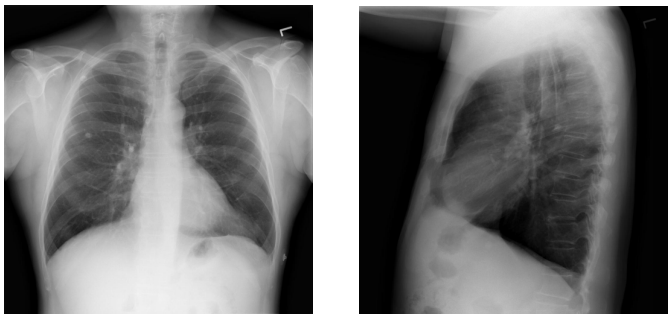


FIGURE 1: EXAMPLE CHEST X-RAY VIEWS USED IN CLINICAL REPORTING: (LEFT) FRONTAL VIEW; (RIGHT) LATERAL VIEW.

2.2 Vision-Language Models in Medical Imaging

Vision-language models (VLMs) have emerged as powerful tools for multimodal learning, capable of aligning image features with textual descriptions. MedCLIP [4] and BioViL [5] enhanced cross-modal representation learning in the medical domain by incorporating biomedical corpora into contrastive training. However, most VLMs are optimized for retrieval or classification tasks and have limited ability to generate coherent, structured diagnostic narratives—a key requirement in clinical settings.

2.3 Radiology Report Generation

Report generation models such as TieNet [6] and R2Gen [7] mark a shift from fixed-label predictions to free-form report synthesis. These systems integrate image-text attention mechanisms or memory modules to improve fluency and coherence. Nonetheless, challenges remain: hallucinated findings, lack of factual grounding, and weak interpretability continue to affect their clinical reliability.

2.4 Limitations in Existing Work

Many existing systems rely on single-view chest X-rays, despite multi-view imaging being standard practice for radiologists. Additionally, most models are trained on relatively small paired datasets (e.g., OpenI), limiting their generalizability across varied radiographic presentations. Furthermore, few efforts explicitly generate reasoning traces to justify diagnoses—an important aspect for educational and clinical auditability.

2.5 Our Contribution in Context

To address these gaps, our work introduces several innovations:

1. Multi-view support to better align with real-world diagnostic workflows.

2. Hybrid dataset training combining expert-written (OpenI) and synthetically generated (CheXpert) reports for broader exposure.

3. Reinforcement learning optimization to improve factual accuracy and include reasoning explanations alongside findings.

TABLE 1: COMPARISON OF OUR SYSTEM WITH PRIOR CHEST X-RAY INTERPRETATION MODELS

Feature	**CheXNet** [3]	**R2Gen** [7]	**BioViL** [5]	**Our System**
Task	Disease Classification	Report Generation	VL Pretraining	Report + Reasoning
Model Type	CNN	Transformer	VL Encoder	VL Transformer (Qwen [9] 2.5B)
Input	Single CXR	Single CXR	Single CXR	Multi-view CXR
Output	Fixed Labels	Free-form Text	Embeddings	Structured Report + Reasoning
Data Used	ChestX-ray14	MIMIC-CXR	PubMed + CXR	OpenI + Synth. CheXpert
Explainability	No	No	No	RL Traces
Factual Accuracy	No	No	No	RL Reward
Efficiency	No	No	No	LoRA + GRPO
Deployable	No	No	No	Modular + Lightweight

4. Efficient fine-tuning techniques (LoRA [10] + GRPO [11]) to enable scalable training of large models under hardware constraints.

In contrast to prior systems, our approach focuses not only on linguistic fluency but also on interpretability and clinical relevance, positioning it as a promising foundation for radiology education and decision support.

Table 1 compares our system with representative prior models, highlighting improvements in interpretability, factual grounding, and real-world deployability.

3. PROBLEM STATEMENT AND PROJECT GOALS

3.1 Motivation

Despite significant progress in automated chest X-ray interpretation, current AI systems predominantly rely on classification models that output fixed disease labels. These models, while useful for detection, lack the narrative depth, diagnostic reasoning, and clinical context found in radiologist-written reports. This limitation hinders their usability in real-world workflows and offers minimal value in educational contexts.

Moreover, interpretability remains a critical barrier. Most existing systems cannot explain *why* a diagnosis was made, making them unsuitable for training purposes or clinical environments that demand auditability. Data scarcity and resource constraints further complicate efforts to train large, expressive models tailored for radiology report generation.

There is a clear need for AI systems that not only identify pathologies in chest X-rays but also generate structured, explanatory reports that mirror radiologist thinking—particularly in environments where expert feedback is limited or unavailable.

3.2 Objectives

The goal of this project is to design and implement an interpretable, resource-efficient AI system that generates structured radiology reports from chest X-ray images using a vision-language modeling approach. The system aims to support both diagnostic assistance and radiology education.

To achieve this, we define the following objectives:

1. **Develop a Vision-Language Model (VLM)**

   Fine-tune a multimodal transformer (Qwen 2.5 7B) capable of generating free-form, radiologist-style reports from single or multi-view chest X-rays.

2. **Leverage Hybrid Training Data**

   Combine high-quality, expert-authored reports from the OpenI dataset with synthetically generated reports for CheXpert images to improve generalization and dataset scale.

3. **Ensure Factual Accuracy and Interpretability**

   Introduce a reinforcement learning module that rewards alignment between generated reports and visual evidence, encouraging the model to produce explanatory reasoning alongside findings.

4. **Optimize for Hardware Constraints**

   Use Low-Rank Adaptation (LoRA) and Gradient Reprocessing Optimization (GRPO) to enable fine-tuning of large models on modest compute infrastructure.

5. **Evaluate Linguistic and Clinical Validity**

   Evaluate the system's report generation capabilities using natural language generation metrics such as BLEU [13], ROUGE-L [14], and Exact Match to assess fluency, structure, and accuracy. Additionally, track improvements in factual consistency during the optimization phase using internal performance thresholds to guide training..

4. METHODOLOGY

4.1 Project Evolution and Challenges

The development of the Multimodal Chest Disease Diagnosis System followed an iterative approach shaped by both technical constraints and evolving insights. While our core objective—generating structured, radiologist-style reports from chest X-rays—remained constant, the project underwent key architectural shifts to address limitations in data availability and compute resources.

We began with a supervised learning approach using the OpenI dataset, which includes paired chest X-rays and expert-written reports. Our initial model selection focused on Qwen-VL 3B, a lightweight vision-language transformer. However, two major challenges emerged early:

1. Hardware limitations: The model struggled with long-sequence inputs, causing inference delays and memory overflows on available GPUs.

2. Limited dataset diversity: OpenI's small scale and lack of rare pathologies restricted generalizability, especially in clinically nuanced cases.

To validate our preprocessing pipeline and establish a baseline, we temporarily pivoted to a CNN-based multi-label classifier trained on the CheXpert dataset. This model was effective for label prediction but fell short in delivering narrative-style outputs, interpretability, and diagnostic depth. It reinforced the need to return to a generative vision-language modeling framework.

In the next phase, we transitioned to Qwen 2.5 7B, an open-source large vision-language model better suited for free-form report generation. To expand training coverage, we generated synthetic reports for CheXpert images using a pretrained medical report generator, mimicking the OpenI structure. This hybrid dataset significantly improved scale and variability, enabling the model to learn from both real and synthetic clinical narratives.

To enable training within hardware constraints, we adopted Low-Rank Adaptation (LoRA) for efficient fine-tuning and later integrated a second-stage optimization technique to improve factual alignment.

This adaptive development process allowed us to build a scalable, interpretable system while managing real-world limitations in data volume, computational power, and clinical complexity.

4.2 System Architecture

The final system was designed to generate high-quality, structured radiology reports from chest X-ray images while remaining interpretable and computationally efficient. To achieve this, we implemented a modular, multi-stage architecture that integrates:

- Vision-language modeling for image-to-text generation
- Hybrid dataset training combining real and synthetic reports
- Efficiency-focused fine-tuning strategies (e.g., LoRA [10])
- Follow-up optimization phase to improve factual grounding

At a high level, the architecture consists of:

1. **Data Pipeline:**

   A hybrid dataset composed of expert-authored (OpenI) and synthetically generated (CheXpert) radiology reports was curated. Preprocessing steps ensured data quality, structural consistency, and balance across pathologies.

2. **Model Backbone:**

   We adopted Qwen 2.5 7B, an autoregressive transformer model capable of processing both image and text inputs. A frozen vision encoder extracts latent features from the X-ray, which are passed to a language decoder conditioned on structured prompts.

3. **Training Workflow:**

   The training was conducted in two stages. First, the model was fine-tuned on the hybrid dataset using supervised learning with cross-entropy loss. To reduce computational overhead, we applied Low-Rank Adaptation (LoRA), updating only a subset of trainable parameters.

4. **Optimization for Factual Consistency:**

   In the second phase, the model was further refined using an optimization strategy that encouraged alignment between generated content and visible image features. This stage introduced structured output formats (e.g., <report>, <think>) to support both diagnostic statements and explanatory reasoning.

5. **Inference Design:**

   The system was configured for lightweight inference, producing complete reports—including findings, impressions, and reasoning traces—within seconds on consumer-grade GPUs. Output formats were designed to integrate with educational platforms or clinical prototypes via JSON or plain text.

4.2.1 Data Sources and Pre-processing. To train our report generation model, we constructed a hybrid dataset that combined real expert-authored reports with synthetically generated ones. This approach aimed to maximize both data quality and scale, which are essential for training large vision-language models in the medical domain.

Data Sources: We utilized two major public datasets:

Table 2 summarizes the two datasets used in this project. While OpenI offers high-quality supervision through paired reports, CheXpert provides the scale needed for robust training after augmentation via synthetic text generation.

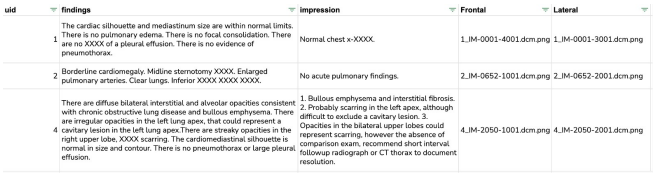
TABLE 2: COMPARISON OF OPENI AND CHEXPERT DATASETS

Dataset	Source	Content	Relevance
OpenI	National Library of Medicine	3,955 radiology reports and 7,470 chest X-ray images (1998–2019)	Enables text-image correlation for multimodal report generation and language grounding
CheXpert	Stanford University	224,316 chest X-rays from 65,240 patients labeled for 14 thoracic conditions (Oct 2002 – Jul 2017)	Robust source for pretraining and generating synthetic paired reports using medical language models

While CheXpert provides structured diagnostic labels for over 220,000 chest X-rays, it does not contain full narrative radiology reports. To enable natural language training on this dataset, we used **CXR-LLaVA**, a pretrained multimodal vision-language model, to generate synthetic reports for selected CheXpert images. These generated reports were structured to match the OpenI format, with `<Findings>` and `<Impression>` sections, ensuring consistency in both content and annotation style. This process significantly increased the volume and variability of training data, exposing the model to a broader spectrum of pathologies, linguistic patterns, and imaging conditions. The augmented dataset—comprising OpenI and synthetic CheXpert samples—was then passed through a unified preprocessing pipeline prior to data splitting.

Preprocessing Steps: To ensure data consistency, interpretability, and model compatibility, we applied the following preprocessing steps:

- *Image Preparation:*



uid	findings	impression	Frontal	Lateral
1	The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.	Normal chest x-XXXX.	1_IM-0001-4001.dcm.png	1_IM-0001-3001.dcm.png
2	Borderline cardiomegaly. Midline sternotomy XXXX. Enlarged pulmonary arteries. Clear lungs. Inferior XXXX XXXX XXXX.	No acute pulmonary findings.	2_IM-0652-1001.dcm.png	2_IM-0652-2001.dcm.png
4	There are diffuse bilateral interstitial and alveolar opacities consistent with chronic obstructive lung disease and bullous emphysema. There are irregular opacities in the left lung apex, that could represent a cavitary lesion in the left lung apex.There are streaky opacities in the right upper lobe. XXXX scarring. The cardiomediastinal silhouette is normal in size and contour. There is no pneumothorax or large pleural effusion.	1. Bullous emphysema and interstitial fibrosis. 2. Probably scarring in the left apex, although difficult to exclude a cavitary lesion. 3. Opacities in the bilateral upper lobes could represent scarring, however the absence of comparison exam, recommend short interval followup radiograph or CT thorax to document resolution.	4_IM-2050-1001.dcm.png	4_IM-2050-2001.dcm.png

FIGURE 2: SNIPPET OF OPENI DATASET

  All chest X-ray images were resized to 224×224 pixels and normalized to match the input requirements of the vision encoder.

- *Text Filtering and Parsing:*

  From OpenI reports, only the "Findings" and "Impression" sections were retained. Synthetic CheXpert reports were filtered to remove incoherent or hallucinated content and truncated to fit the model's context window.

- *Class-Balanced Sampling:*

  A custom sampling strategy was implemented to prevent overrepresentation of "No Finding" cases. This ensured the model encountered a balanced variety of pathologies during training.

Data Splitting and Balance: From a combined pool of 2,200 image-report pairs (both real and synthetic), we created a stratified data split: 1,540 samples for training; 220 samples for validation; 440 samples for testing

This split preserved class distribution across pathology types and maintained variation in report length and diagnostic complexity. The validation and test sets were held out entirely during training and used exclusively for model selection and evaluation.

4.2.2 Model Selection and Fine-Tuning. The core of our system is Qwen 2.5 7B, an open-source, autoregressive vision-language transformer capable of generating structured text based on multimodal input. Compared to smaller vision-language models used in our early experiments, Qwen 2.5 7B demonstrated improved capacity to model long sequences, better handling of medical terminology, and more coherent diagnostic text generation.

Model Architecture The architecture of Qwen 2.5 7B comprises two main components:

- A frozen vision encoder that processes chest X-ray images into high-dimensional embeddings. This encoder extracts visual tokens that encapsulate radiographic features.

- A language decoder that autoregressively generates diagnostic reports, conditioned on both visual tokens and an input text prompt.

The input pipeline begins by embedding the image and appending a structured prompt that instructs the model to generate a radiology report. A typical input prompt is:

```
<prompt> Generate diagnostic report for the
following chest X-ray: <image> </prompt>
<report>
  <findings> ... </findings>
  <impression> ... </impression>
</report>
```

This prompt format was standardized across all samples to encourage the model to adhere to structured output conventions and clinical tone.

Fine-Tuning Strategy Due to the size of the model and our limited hardware availability, we adopted a parameter-efficient fine-tuning strategy to train the model effectively without overloading GPU memory. The fine-tuning process was carried out in two stages:

1. **Supervised Fine-Tuning (SFT) (Phase 1):**

   In the first phase, we fine-tuned the model on our hybrid dataset (OpenI + synthetic CheXpert reports) using cross-entropy loss, which measured the token-level divergence between predicted outputs and ground truth sequences.

   To enable training on limited hardware (four NVIDIA H100 GPUs), we implemented the following optimizations:

   - *Low-Rank Adaptation (LoRA):*
     LoRA introduced lightweight trainable matrices into the attention layers of the transformer. This allowed us to update a small subset of parameters while keeping the rest of the model frozen. It significantly reduced the memory footprint without compromising learning capacity.
   - *Gradient Reprocessed Policy Optimization (GRPO):*
     GRPO was used to improve training stability and gradient flow, particularly in low-resource environments. Although originally designed for policy optimization, we adapted GRPO to maintain steady updates under limited memory and noisy gradient conditions.

   This phase trained the model to reproduce expert-authored reports in a structured format, learning medical language patterns and disease-specific phrasing effectively.

2. **Preparation for Reinforcement Learning (Phase 2):**

   The outputs from this stage were already linguistically coherent and structurally aligned with clinical report conventions. However, the model occasionally exhibited hallucinations—generating plausible but incorrect findings not supported by the image.

   To address this, we prepared the model for a second training stage (described in Section 4.2.3) using reinforcement learning guided by a custom reward function. This transition allowed us to refine factual consistency and interpretability by encouraging the generation of explanatory reasoning along with the report.

By the end of the fine-tuning phase, the model could generate full diagnostic narratives—including <findings> and <impression> sections—with strong stylistic similarity to real radiologist reports. This laid the foundation for the next phase, where factual alignment and clinical justification were further optimized using reinforcement learning.

4.2.3 Reinforcement Learning for Factual Consistency. While the fine-tuned Qwen 2.5 7B model exhibited strong language fluency and report structure, it occasionally produced hallucinated outputs—statements that were medically plausible but not grounded in the actual chest X-ray. To address this, we implemented a second optimization stage using Reinforcement Learning (RL) to explicitly reward factual accuracy and structured reasoning.

This RL phase aimed to improve the alignment between generated diagnostic reports and the visual features of the X-ray, while also encouraging interpretable clinical justifications.

Dual-Output Generation Objective: During reinforcement learning, the model was prompted to generate two distinct but related outputs:

- A structured diagnostic report, enclosed within <report>, <findings>, and <impression> tags
- A corresponding reasoning trace, enclosed within a <think> tag

This output format reflects how radiologists not only document observations but also reason through the image-to-diagnosis process. For example:

```
<report>
<findings> Right costophrenic angle blunting
</findings>
<impression> Suggests mild pleural effusion
</impression>
</report>
<think> Blunting of the right costophrenic angle
with a meniscus shape indicates fluid
accumulation.
</think>
```

Reward Function Design and Implementation: To shape the model's behavior during training, we implemented a custom reward function called accuracy_reward(), which provides structured feedback for both report content and reasoning accuracy.

Reward computation follows four key stages:

1. Tag-Based Matching: For each generated sample, the function extracts content within <report> and <think> tags and compares it to the ground truth using string similarity.
2. Reward Scoring:
   - 0.5 reward is granted for an exact match in the diagnostic report (<report>)
   - +0.5 reward for a correct explanation (<think>)

- Total reward ranges from 0.0 to 1.0 per sample

3. Fallback Logic: If tags are missing or malformed, the function falls back to raw string comparisons to maintain training robustness.

4. Debug Logging: When enabled, DEBUG_MODE logs reward scores, mismatches, and timestamps for inspection. This helped refine prompt strategies and detect systematic errors during training.

This reward formulation allowed the model to receive fine-grained feedback, encouraging it to internalize both clinical knowledge and structured reporting patterns.

Optimization with GRPO: We used Gradient Reprocessed Policy Optimization (GRPO) as the RL algorithm. GRPO is a lightweight, resource-efficient policy gradient method that supports large-model fine-tuning with improved gradient flow and reduced variance. It was particularly well-suited to our memory-constrained setup.

Training Configuration: Due to the scale of Qwen 2.5 7B and the overhead of dual-output supervision, we trained with:

- *Batch size: 8*
- *Gradient accumulation: Enabled*
- *CPU offloading: To reduce VRAM usage*

Training was halted once the average reward-based loss converged below a defined threshold.

Post-RL Improvements: Reinforcement learning yielded clear performance improvements:

- Impression sections became more concise and evidence-grounded
- Reasoning traces more reliably cited visible image features
- Hallucinations were reduced in favor of precise, image-supported findings

For example, vague outputs like "possible fluid buildup" evolved into specific statements such as:

"Blunting of the right costophrenic angle indicates mild pleural effusion."

This behavior suggests improved diagnostic accountability, bridging the gap between language generation and clinical reasoning.

4.2.4 Inference and Output Format. After training, the Qwen 2.5 7B model was deployed in an inference pipeline designed to generate diagnostic reports for new chest X-ray images in real-time. This phase focused on delivering clinically useful, structured outputs while maintaining low latency and memory efficiency.

Inference Pipeline: The inference process consists of the following stages:

1. Image Encoding

   The input chest X-ray is resized and normalized, then passed through the frozen vision encoder, which produces a fixed-length sequence of visual embeddings.

2. Prompt Injection

   A predefined prompt, containing structural tags such as <report>, <findings>, <impression>, and optionally <think>, is prepended to the encoded image features. This template guides the model in generating structured outputs.

3. Autoregressive Decoding

   The language decoder generates tokens one-by-one, conditioned on both the visual embeddings and the prompt. Beam search decoding is used to improve fluency and consistency in output.

4. Post-Processing

   The generated sequence is parsed to extract content within the defined tags. If any tags are missing or malformed, fallback logic is applied to extract the most relevant text segments.

Performance and Resource Profile: Inference was tested on a Quad-GPU system (NVIDIA H100), with each chest X-ray processed in approximately 3–6 seconds, depending on sequence length and decoding configuration. The use of a frozen vision encoder and LoRA-enabled decoder allowed for efficient runtime memory usage, making the system suitable for batch processing and interactive educational tools.

This modular and lightweight inference setup supports practical deployment scenarios — from radiologist-assistive tools to AI-augmented learning platforms.

4.2.5 Evaluation Strategy. We adopted a two-pronged evaluation strategy aligned with the system's training phases: (1) supervised fine-tuning and (2) reinforcement learning. Each phase used metrics appropriate to its learning objective—linguistic quality for fine-tuning and factual alignment for reinforcement optimization.

Evaluation After Supervised Fine-Tuning: To assess the quality of the model's outputs after fine-tuning on the hybrid dataset, we employed widely accepted natural language generation (NLG) metrics:

- BLEU (Bilingual Evaluation Understudy):

  Measures n-gram overlap between the generated and reference texts. High BLEU scores indicate strong phrase-level similarity and fluency.

- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation):

  Focuses on the longest common subsequence between generated and ground-truth reports. It reflects how well the model retains key diagnostic elements and narrative structure.

- Exact Match (EM):

  Captures the percentage of generated reports that exactly match the reference outputs in both content and format. Although strict, this metric benchmarks perfect reproduction of expert-authored reports.

These metrics were calculated on the held-out test set of 440 samples, ensuring that the evaluation remained independent of training and validation data.

Evaluation After Reinforcement Learning: Since reinforcement learning was designed to optimize factual consistency and clinical reasoning, traditional NLG metrics were insufficient to capture improvement. Therefore, we introduced a custom reward-based evaluation, aligned with the accuracy_reward() function used during training.

- Reward Function Overview:

  For each sample, the model's output was parsed into <report> and <think> segments. The reward function assigned points based on function mentioned earlier.

- Loss Monitoring and Convergence:

  Reinforcement training was monitored using the reward-converted loss value. Training was halted when this loss fell below a predefined threshold, indicating stable, high-quality output generation.

- Qualitative Assessment:

  Manual audits of post-RL outputs showed improved alignment between impressions and visual features, reduced hallucinations, and more concise reasoning. For example:

This two-stage evaluation strategy ensured that the system was not only linguistically fluent, but also clinically accurate and interpretable, aligning with the broader goals of radiology automation and education.

5. RESULTS AND EXAMPLES

We present the results of our system in two parts: quantitative performance based on evaluation metrics, and qualitative examples that illustrate improvements in factual grounding, fluency, and clinical reasoning.

5.1 Quantitative Results

The table below summarizes model performance before and after reinforcement learning, based on the test set of 440 samples:

- BLEU and ROUGE-L showed minor but consistent improvements, reflecting better linguistic and structural fidelity.
- Exact Match increased, indicating cleaner formatting and improved tag compliance.
- Reward Score, introduced in the RL phase, captured semantic and structural accuracy—confirming factual consistency and improved reasoning.

TABLE 3: MODEL PERFORMANCE COMPARISON

Metric	Fine-Tuned	Post-RL
BLEU Score	0.83	0.84
ROUGE-L Score	0.93	0.94
Exact Match	0.75	0.77
Reward Score	—	0.91

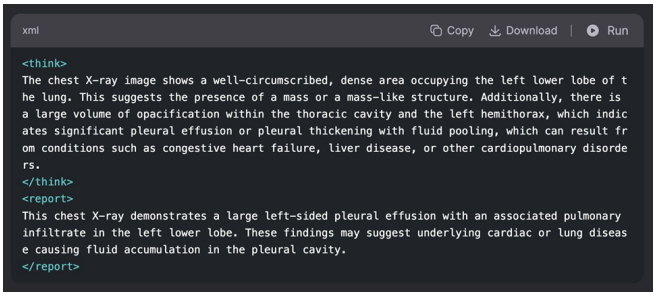


FIGURE 3: SAMPLE OUTPUT OF MODEL

5.2 Qualitative Examples

Figure 3 presents a sample radiology report generated by the model, while Figure 4 provides a comparative view, displaying the model's output alongside the ground truth findings for reference.

6. CONCLUSION AND FUTURE WORK

This project presents a multimodal AI system capable of generating structured, interpretable radiology reports directly from chest X-ray images. By leveraging a vision-language transformer (Qwen 2.5 7B), a hybrid dataset combining expert-authored and synthetic reports, and a two-phase training strategy—supervised fine-tuning followed by reinforcement learning—we developed a system that balances linguistic fluency, clinical accuracy, and diagnostic reasoning.

The system was designed with efficiency and interpretability in mind. Low-Rank Adaptation (LoRA) enabled memory-efficient fine-tuning on limited hardware, while Gradient Reprocessed Policy Optimization (GRPO) guided the model toward factual alignment. The inclusion of structured prompts and reasoning traces (<report>, <findings>, <impression>, <think>) enabled outputs suitable for both diagnostic workflows and radiology education.

Evaluation results demonstrated strong performance across BLEU (0.84), ROUGE-L (0.94), and reward-based metrics (0.91), with qualitative examples showing improved reasoning and reduced hallucination after reinforcement learning. However, several limitations were observed: the model remained sensitive to prompt phrasing, and training required careful optimization under resource constraints. Additionally, while synthetic reports from CXR-LLaVA expanded coverage, they introduced stylistic differences that occasionally affected consistency.

Despite these challenges, the system highlights the feasibility of applying large-scale vision-language models to clinically

<think>
The chest X-ray shows evidence of a pneumothorax on the left side. The lung fields exhibit hyperinflation, with the lung markings being less distinct. There is a large area of opacification on the left side, indicating the presence of air between the pleural surfaces, which is the hallmark of a pneumothorax. Additionally, the ribs on the left side appear hyperlucent (less dense), possibly due to the presence of air. The heart borders and diaphragm shapes appear normal without overt signs of cardiomegaly or diaphragmatic movement that would suggest a pneumothorax.
</think>

<report>
There is a significant pneumothorax on the left side of the chest, with hyperinflation and hyperlucency of the left lung field. The ribs on the left side appear more hyperlucent than on the right. The right lung field and heart borders are normal. There does not appear to be signs of cardiomegaly or other significant pathologies that might explain this finding other than the pneumothorax.
</report>

Solution:
<think> Negative. **</think>
<report>The cardiomediastinal silhouette is normal in size and contour. No focal consolidation, pneumothorax or large pleural effusion. Negative for acute bone abnormality.</report>

FIGURE 4: SAMPLE OUTPUT OF MODEL AND GROUND TRUTH

grounded, explainable report generation. With continued development, it holds potential for use in both real-world radiology assistance and trainee education.

Future work can focus on several directions:

- **Expert feedback loops:** Human-in-the-loop training with radiologists could improve sensitivity to rare findings and enhance reward function design.
- **Multi-modal context integration:** Incorporating patient history, clinical notes, or lab results could help contextualize findings and enhance diagnostic precision.
- **Real-world deployment:** UI/UX design, API standardization, and regulatory compliance will be key for integrating the system into clinical or educational environments.
- **Dataset and modality expansion:** Evaluating the model on larger and more diverse datasets (e.g., MIMIC-CXR, ChestX-ray14) and adapting it to modalities like CT or MRI would broaden applicability.
- **Bias and uncertainty detection:** Mechanisms for estimating model confidence and detecting out-of-distribution inputs would improve safety and transparency.
- **Educational interfaces:** A visual platform for residents to explore model outputs, view reasoning, and receive feedback could support radiology learning environments.

In summary, this project shows how a well-optimized vision-language AI system, rooted in domain expertise and technical rigor, can support medical education and diagnostic workflows. With further refinement, such systems could become reliable collaborators in radiology and useful tools for medical training.

REFERENCES

[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M., 2017. "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106.

[2] Lakhani, P., and Sundaram, B., 2017, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," *Radiology*, 284(2), pp. 574–582.

[3] Rajpurkar, P., et al., 2017, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*.

[4] Wang, H., Xia, Y., and Song, S., 2022, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," *arXiv preprint arXiv:2210.10163*.

[5] Boecking, B., Uszkoreit, J., Boley, M., et al., 2022. "Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing," *Nature Communications*, 13(1), pp. 1–12.

[6] Wang, X., et al., 2018, "TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 9049–9058.

[7] Chen, M., et al., 2020, "Generating Radiology Reports via Memory-Driven Transformer," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 522–532.

[8] Irvin, J., et al., 2019. "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *Proc. AAAI Conf. Artif. Intell.*, 33(01), pp. 590–597.

[9] Qwen Team, 2024, "Qwen-VL: A Vision-Language Model for Multi-Modal Tasks," *https://github.com/QwenLM/Qwen-VL*, Accessed Apr. 2025.

[10] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Wang, W., and Chen, W., 2022, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*.

[11] He, Z., Zhang, Q., Li, D., Wang, C., and Liu, T., 2024, "GRPO: Gradient Reprocessing Optimizer for Fine-Tuning LLMs Under Low-Resource Constraints," *arXiv preprint arXiv:2402.03300*.

[12] Li, J., Song, T., Wang, Z., Yao, Z., Zheng, H., Xu, Y., Zhang, M., Ma, K., and Zheng, Y., 2024, "InstructRadiology: Empowering Vision-Language Models for Medical Report Generation Through Instruction Tuning," *arXiv preprint arXiv:2406.14117*.

[13] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., 2002, "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.

[14] Lin, C.-Y., 2004, "ROUGE: A Package for Automatic Evaluation of Summaries," *Text Summarization Branches Out: Proc. ACL-04 Workshop*, pp. 74–81.