Capstone Project- Final Report
Dual View SLaVA CXR: Enhancing Chest X-Ray Diagnosis

Team 15 | Clinton K J, Rahul Ramesh O, Farnaz Bigdeli, Kyatham Hemanth

1. Abstract:

Chest X-rays (CXRs) are among the most used diagnostic imaging tools in healthcare. However, most AI models, including the SLaVA-CXR model, rely on a single frontal view, limiting diagnostic accuracy. Our capstone project proposes Dual View SLaVA-CXR, which integrates both frontal and lateral views to better emulate clinical radiology workflows. Built upon the Re^3 (Recognize, Reason, Report) paradigm, our lightweight solution extends SLaVA-CXR by incorporating dual-view inputs, enhancing anatomical cross-referencing and clinical reasoning. We use the MIMIC-CXR dataset with over 30,000 dual-view studies, structured reports, and novel vision-language fusion using LLaVA-Phi. Results show significant improvements across BLEU, ROUGE, METEOR, BERTScore, RadGraph F1 and CheXbert F1 scores compared to single-view models, demonstrating the feasibility of Dual-view CXR analysis for real-world deployments.

2. Introduction:

Medical imaging is crucial for diagnosing thoracic diseases. Traditional AI models for chest X-ray interpretation are limited to single views, often ignoring the lateral projection which is vital for detecting abnormalities like hidden masses or effusions. Our project bridges this gap by extending the SLaVA-CXR model to include both frontal and lateral views, improving the quality of automated radiology report generation. This mirrors how radiologists operate in real-world settings.

3. Literature Review:

SLaVA-CXR (Base Paper) [1] uses a lightweight Re^3 pipeline to generate reports from single-view CXRs, but lacks dual-view integration, limiting clinical relevance.

PromptMRG [2] improves report generation via diagnosis-aware prompts and adaptive losses but lacks multi-view support and fact verification.

PMC-VQA [3] handles visual QA using medical data but doesn't address report generation or dual-view reasoning.

CheXFusion [4] fuses dual-view CXRs for classification using cross-attention and a frozen backbone, excelling on long-tailed diseases but is limited to classification.

While these models contribute unique innovations but none of them combine dual-view imaging with lightweight, structured report generation. Our model addresses this gap by integrating LLaVA-Phi with dual-view fusion for anatomically grounded reports.

Limitations in prior work:

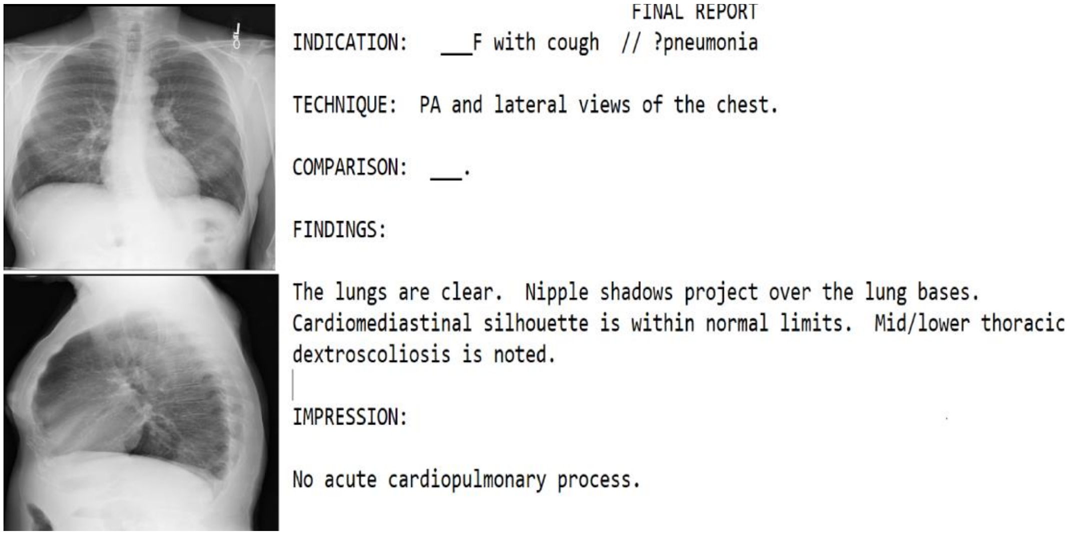
- No support for multi-view CXR inputs
- No structured report generation
- No factual consistency verification
- Poor adaptability to real-world clinical data

4. Materials and Method

4.1 Materials

Dataset and Source

We used the MIMIC-CXR v2.0.0 dataset from PhysioNet, a large-scale, de-identified chest X-ray repository for academic use. It contains over 377,000 images from 227,827 studies with corresponding radiology reports, including frontal and lateral views. Below shows a single study sample Fig 1.



```
                     FINAL REPORT
INDICATION:   ___F with cough  // ?pneumonia

TECHNIQUE:  PA and lateral views of the chest.

COMPARISON:  ___.

FINDINGS:

The lungs are clear.  Nipple shadows project over the lung bases.
Cardiomediastinal silhouette is within normal limits.  Mid/lower thoracic
dextroscoliosis is noted.

IMPRESSION:

No acute cardiopulmonary process.
```

Figure 1: Sample frontal and lateral X-ray images of MIMIC-CXR data set with report

Selection Criteria

To support dual-view diagnostic learning, we selected studies with exactly one frontal and one lateral image, and complete Findings + Impression sections in the report. After filtering for missing views or corrupted data, we curated a final working set of ~30,000 dual-view studies to match resource limits.

4.2 Method

This work extends the SLaVA-CXR framework by integrating dual-view chest X-ray analysis into the LLaVA-Phi architecture. Our methodology follows a modified Recognize–Reason–Report (RE³) paradigm which is illustrated below:

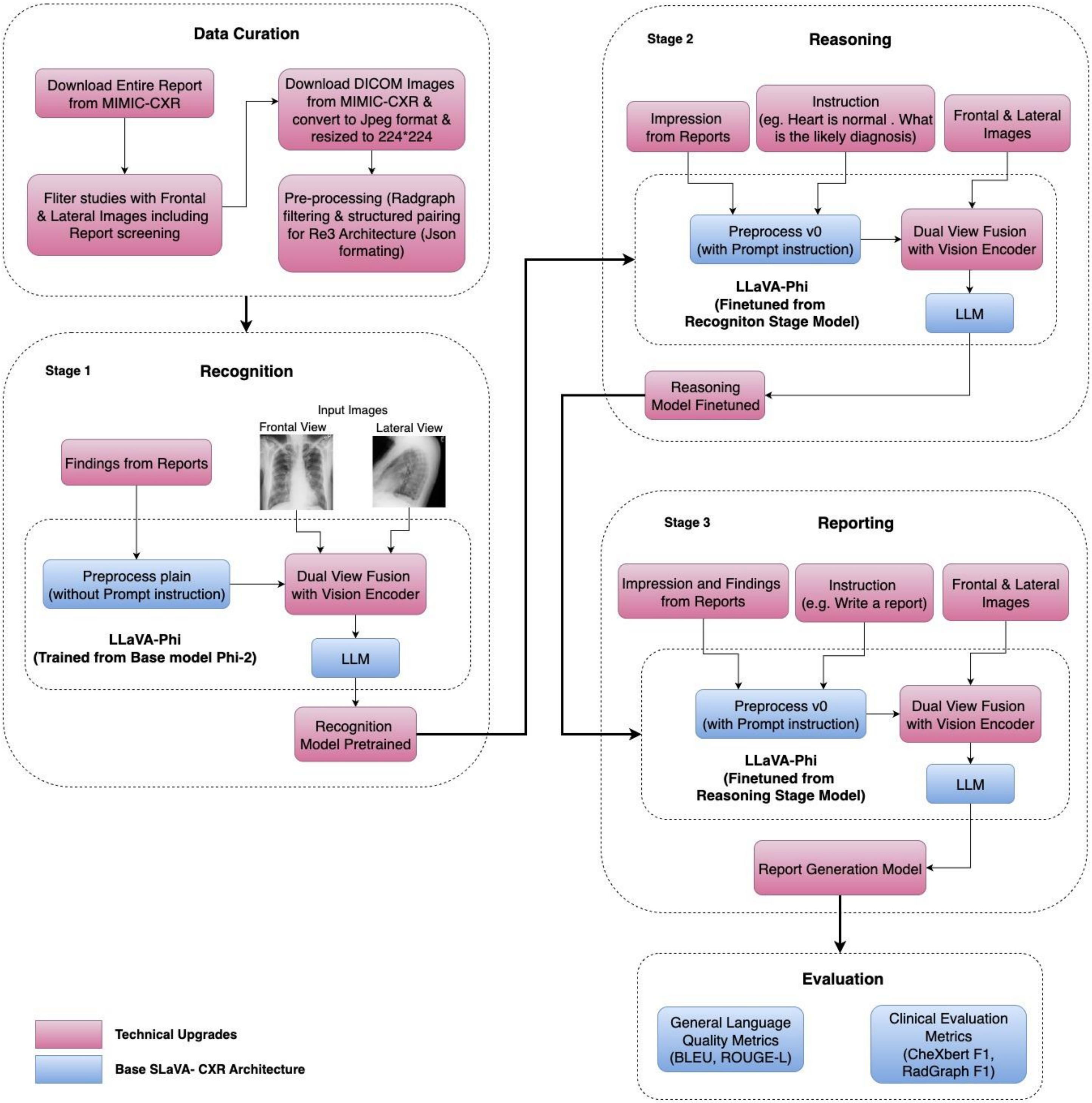

Data Curation
Download Entire Report from MIMIC-CXR
Download DICOM Images from MIMIC-CXR & convert to Jpeg format & resized to 224*224
Fliter studies with Frontal & Lateral Images including Report screening
Pre-processing (Radgraph filtering & structured pairing for Re3 Architecture (Json formating)
Stage 1
Recognition
Input Images
Frontal View
Lateral View
Findings from Reports
Preprocess plain (without Prompt instruction)
Dual View Fusion with Vision Encoder
LLaVA-Phi (Trained from Base model Phi-2)
LLM
Recognition Model Pretrained
Stage 2
Reasoning
Impression from Reports
Instruction (eg. Heart is normal . What is the likely diagnosis)
Frontal & Lateral Images
Preprocess v0 (with Prompt instruction)
Dual View Fusion with Vision Encoder
LLaVA-Phi (Finetuned from Recogniton Stage Model)
LLM
Reasoning Model Finetuned
Stage 3
Reporting
Impression and Findings from Reports
Instruction (e.g. Write a report)
Frontal & Lateral Images
Preprocess v0 (with Prompt instruction)
Dual View Fusion with Vision Encoder
LLaVA-Phi (Finetuned from Reasoning Stage Model)
LLM
Report Generation Model
Evaluation
General Language Quality Metrics (BLEU, ROUGE-L)
Clinical Evaluation Metrics (CheXbert F1, RadGraph F1)
Technical Upgrades
Base SLaVA- CXR Architecture

Figure 2: Re^3 Training Pipeline for Dual-View SLaVA-CXR. The model progresses through recognition, reasoning, and reporting stages using dual-view chest X-rays, integrating Dual-View fusion and prompt-based supervision for structured radiology report generation.

Data Curation

All DICOM images were converted to JPEG and resized to 224×224 pixels. Matched image–report pairs were structured into JSON format. We applied RadGraph parsing to extract structured observation–anatomy triplets and filter out low-quality text content (see Appendix 10.3).

Dual-View Fusion (Architectural Contribution)

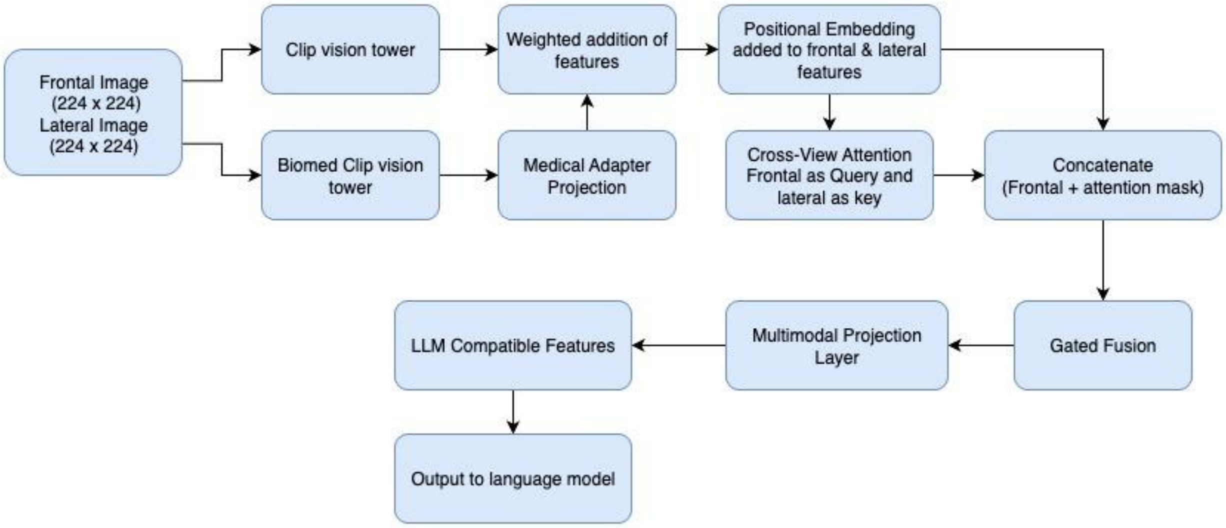

Frontal Image (224 x 224) Lateral Image (224 x 224)
Clip vision tower
Biomed Clip vision tower
Weighted addition of features
Medical Adapter Projection
Positional Embedding added to frontal & lateral features
Cross-View Attention Frontal as Query and lateral as key
Concatenate (Frontal + attention mask)
Gated Fusion
Multimodal Projection Layer
LLM Compatible Features
Output to language model

Figure 3: Dual-view fusion pipeline for radiology reporting. Frontal and lateral X-rays are encoded via CLIP and BiomedCLIP, with a learnable adapter and cross attention. Gated fusion produces LLM-compatible features for structured report generation

To support clinical reasoning from frontal and lateral chest X-rays, we introduce a dual-view fusion module (see Fig 3) that processes each view using two frozen vision encoders: CLIP for general semantics and BiomedCLIP for medical specificity. BiomedCLIP features are passed through a linear adapter and layer normalization to align with CLIP's feature space.

Let f_{CLIP} and f_{Biomed} be the features from CLIP and BiomedCLIP respectively. A learnable scalar weight $\alpha = \sigma(w)$ determines their contribution, and the aligned features are fused as:

$$f_{view} = \alpha \cdot f_{CLIP} + (1 - \alpha) \cdot Adapter(f_{Biomed}), for\ view \in \{frontal, lateral\} \qquad (1)$$

Next, we add learned positional embeddings and apply layer normalization to each view's fused features. To capture anatomical correspondences, we apply cross-attention using frontal features as queries and lateral features as keys and values:

$$a = \mathrm{CrossAttention}(f_{frontal}, f_{lateral}, f_{lateral}) \qquad (2)$$

The frontal features and attention output are concatenated and passed through a gated fusion layer, which computes a dynamic gating value $\beta \in [0,1]$ via an MLP followed by a sigmoid. The final fused representation is:

$$f_{final} = \beta \cdot f_{frontal} + (1 - \beta) \cdot a \tag{3}$$

Finally, this representation is projected into a multimodal embedding space using a learnable projector, producing features compatible with the Phi-2 language model. This architecture allows the model to integrate spatial and semantic cues across views and domains for enhanced diagnostic reasoning and report generation.

Recognize–Reason–Report (Re³) Pipeline

Rather than a single-step model, we adopt a progressive, modular learning strategy that reflects real radiological workflows (see Fig 2):

Stage 1: Recognition

The model learns to generate Findings directly from dual-view inputs without prompts (refer Json). This acts as the base encoder–decoder foundation.

Stage 2: Reasoning

Fine-tuned on top of recognition, this stage uses the generated Findings and a clinical prompt (refer Json) to produce the Impression. This promotes interpretability and diagnostic reasoning.

Stage 3: Reporting

Finally, the model is supervised to generate complete structured reports (Findings + Impression) given both images and a high-level prompt (refer Json).

This Re³ pipeline allows the model to progressively acquire structured clinical understanding while enabling intermediate supervision and targeted improvements. (Training details see Appendix 10.5)

Evaluation

We evaluated model performance using a 95/5 train–test split on MIMIC-CXR (1,596 test samples). The IU X-Ray dataset (2,955 samples) was used exclusively for external evaluation and not included in training.

- Text generation metrics: BLEU, ROUGE-L, METEOR, BERT
- Clinical accuracy: RadGraph F1 (see Radgraph working details) and CheXbert F1 (see CheXbert working details) , which scores both correct clinical entities and relations

Implementation details, system configuration, and hyperparameters are included in Appendix Section 10.4.

Github : Dual View Slava Codes

5. Results:

Table 1: Chest X-ray report generation performance of methods. R-L, M, B-2, BS, CX, RG, and RC are short for ROUGE-L, METEOR, BLEU-2, BERTScore, CheXbert and RadGraph, respectively. All results are reported in percentage (%). ↑ and ↓ denote 'the higher the better' and 'the lower the better', respectively.

Method	MIMIC-CXR						
	Params	**R-L ↑**	**M ↑**	**B-2 ↑**	**BS ↑**	**CX ↑**	**RG ↑**
LLaVA v0 (2023c)	7B	6.90	15.83	2.32	-1.56	32.74	6.14
LLaVA-Med (2023)	7B	5.85	14.31	1.96	1.07	33.92	5.45
LLaVA v1.5 (2023b)	7B	7.87	17.38	2.54	14.26	35.16	6.51
TinyGPT-V (2023)	2.7B	4.48	1.90	0.63	5.04	41.96	0.20
LLaVA-phi (2023c)	2.7B	3.63	13.21	1.17	0.02	32.08	2.42
SLaVA-CXR (Base)	2.7B	9.14	19.92	3.49	20.82	35.24	8.47
Dual View SLaVA (Ours)	**2.7B**	**27.10**	**20.85**	**12.42**	**87.54**	**40.83**	**21.68**

Table 2: Chest X-ray report generation performance of methods. R-L, M, B-2, BS, CX, RG, and RC are short for ROUGE-L, METEOR, BLEU-2, BERTScore, CheXbert and RadGraph, respectively. All results are reported in percentage (%). ↑ and ↓ denote 'the higher the better' and 'the lower the better', respectively

Method	IU-XRay						
	Params	**R-L ↑**	**M ↑**	**B-2 ↑**	**BS ↑**	**CX ↑**	**RG ↑**
LLaVA v0 (2023c)	7B	3.78	10.40	1.12	-6.67	30.90	2.63
LLaVA-Med (2023)	7B	2.86	9.18	0.87	-4.26	31.33	2.28
LLaVA v1.5 (2023b)	7B	1.21	0.72	0.22	-7.51	**71.26**	0.43
TinyGPT-V (2023)	2.7B	1.45	1.22	0.63	-8.53	31.15	2.69
LLaVA-phi (2023c)	2.7B	1.05	5.22	0.21	-10.05	21.58	0.35
SLaVA-CXR (Base)	2.7B	5.08	14.49	3.53	24.17	64.41	3.96
Dual View SLaVA (Ours)	**2.7B**	**19.54**	**19.42**	**8.84**	**85.98**	38.1	**4.83**

Table 3: Qualitative analysis of model outputs. Blue-colored text denotes alignment between the ground truth text and the generated text

Study ID	Reference Report	Model Prediction
51159934	FINDINGS: Lungs are hypoinflated. There is a diffuse bilateral interstitial process, worst at the right lung base. Heart size and mediastinal contours are normal. There is no large pleural effusion. No pneumothorax. Osseous structures appear intact. IMPRESSION: Diffuse interstitial process, concerning for atypical pneumonia, worst at the right lung base.	FINDINGS: The cardiomediastinal and hilar contours are within normal limits. There is no pneumothorax or large pleural effusion. Lungs are well-expanded. Subtle opacity in the right lower lobe is concerning for pneumonia. The upper abdomen is unremarkable. IMPRESSION: Subtle right lower lobe opacity is concerning for pneumonia.
58509000	FINDINGS: Cardiomediastinal contours are stable. Right lung and pleural surfaces remain clear. IMPRESSION: Worsening peripheral left mid-lung consolidation, for which the differential diagnosis includes infectious pneumonia, radiation pneumonitis, organizing pneumonia, and less	FINDINGS: The heart size is normal. The mediastinal and hilar contours are unchanged. Pulmonary vascularity is normal. There is consolidation within the left lower lobe concerning for pneumonia. No pleural effusion or pneumothorax is seen. There are no acute

	likely infarction. RECOMMENDATION(S): Consider chest CT for further characterization of the left lung consolidation.	osseous abnormalities. IMPRESSION: Left lower lobe pneumonia.

Summarization Results

We present chest X-ray report generation results on the MIMIC-CXR and IU-Xray datasets in Tables 1 and Table 2. Despite having fewer parameters than most baselines, our Dual View SLaVA model achieves the best performance across key metrics, with notable gains in ROUGE-L, BERTScore, and RadGraph F1, reflecting improved fluency and clinical accuracy. Compared to LLaVA variants and TinyGPT-V, it benefits from dual-view fusion and a lightweight design. To complement these quantitative results, Table 3 offers qualitative examples showing strong alignment between generated and reference reports, underscoring the model's ability to produce clinically coherent, relevant outputs.

6. Implementation and User Benefit:

Our model integrates a dual-view fusion module into the LLaVA-Phi architecture, enabling analysis of both frontal and lateral chest X-rays. Each view is processed using frozen CLIP and BiomedCLIP encoders, fused via a learnable adapter and attention mechanism. The Recognize–Reason–Report (Re^3) pipeline guides report generation in three stages—Findings, Impression, and Full Report—mimicking real radiological workflows.

This structured, multimodal approach improves the model's ability to detect subtle abnormalities and enhances report accuracy. It benefits radiologists and clinical AI tools by reducing missed diagnoses that commonly occur in single-view models. The system offers faster, anatomically grounded report generation, better aligning with clinical practice and improving decision support without requiring additional imaging or manual annotations.

7. Limitations and Further Improvements:

A key limitation of our work is the limited availability of high-quality dual-view studies in the MIMIC-CXR dataset. Although the dataset includes over 377,000 images, only about 50,000 studies contain both frontal and lateral views with complete Findings and Impression sections. After applying strict quality and abnormality filters, we curated a final dataset of approximately 16,000 abnormal and 14,000 normal studies. This relatively small and imbalanced sample size limits the model's ability to generalize across a wide range of pathologies and affects overall performance.

To address this, future work will focus on expanding the abnormal dataset through improved abnormality detection and enhancing supervision using RadGraph annotations. Our goal is to increase the RadGraph F1 score beyond 0.3 by strengthening factual grounding and structured entity modeling. Additional improvements include cross-domain pretraining, anomaly-aware prompting, and adaptive loss functions to improve clinical correctness, scalability, and diagnostic reliability.

8. Bibliography and References:

1. J. Wu, Y. Kim, D. Shi, D. Cliffton, F. Liu, and H. Wu, "SLaVA-CXR: Small Language and Vision Assistant for Chest X-ray Report Automation," arXiv https://arxiv.org/abs/2409.13321

2. H. Jin, H. Che, Y. Lin, and H. Chen, “PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation,” arXiv (2023). https://arxiv.org/abs/2308.12604

3. X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, “PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering,” arXiv https://arxiv.org/abs/2305.10415

4. D. Kim, “CheXFusion: Effective Fusion of Multi-View Features using Transformers for Long-Tailed Chest X-Ray Classification,” arXiv https://arxiv.org/abs/2308.03968

5. T.-Y. Lin et al., “A Survey of Transformers,” arXiv https://arxiv.org/pdf/2106.13884

6. H. Zeng et al., “Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language,” arXiv https://arxiv.org/pdf/2305.17100

7. P. Wang et al., “Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks,” arXiv https://arxiv.org/pdf/2110.00476

8. A. Vaswani et al., “Attention Is All You Need,” arXiv preprint arXiv https://arxiv.org/pdf/1706.03762

9. S. R. Bowman et al., “A Large Annotated Corpus for Learning Natural Language Inference,” arXiv preprint arXiv https://arxiv.org/pdf/1705.08045

10. Hugging Face, “Transformers: State-of-the-Art Natural Language Processing Library,” https://github.com/huggingface/transformers (accessed July 2025)

11. PyTorch Documentation, “torch.nn.MultiheadAttention,” https://docs.pytorch.org/docs/stable/generated/torch.nn.MultiheadAttention.html (accessed July 2025)

12. S. Zhang, Y. Xu, N. Usuyama et al., “BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs,” arXiv https://arxiv.org/pdf/2303.00915

13. OpenAI, “CLIP: Contrastive Language–Image Pretraining,”

14. https://github.com/openai/CLIP (accessed July 2025)

15. “MIMIC-CXR: A De-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports,” *PhysioNet*, 2019. https://physionet.org/content/mimic-cxr/2.0.0/

16. H. Liu et al., “Visual Instruction Tuning,” arXiv https://arxiv.org/abs/2304.08485

17. S. Bannur et al., “MAIRA-2: Grounded Radiology Report Generation,” arXiv https://arxiv.org/abs/2406.04449

18. Y. Zhu et al., “LLaVA-Phi: Efficient Multimodal Assistant with Small Language Model,” arXiv https://arxiv.org/abs/2401.02330

19. C. Li et al., “LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day,” arXiv https://arxiv.org/abs/2306.00890
20. J. Chen et al., “MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-Task Learning,” arXiv https://arxiv.org/abs/2310.09478

21. A. Smit et al., “CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT,” arXiv preprint arXiv https://arxiv.org/abs/2004.09167

22. CITI, “CITI Data or Specimens Only Research.” https://about.citiprogram.org (accessed July 2025).

23. S. Jain et al., “RadGraph: Extracting Clinical Entities and Relations from Radiology Reports,” arXiv https://arxiv.org/abs/2106.14463

9. Individual Progress:

Each task addresses a specific phase of the workflow, from data preparation to final evaluation and reporting

Task	Sub-task	Owner	ClickUp / Jira Link
Data Filtering	Filtered MIMIC-CXR to retain dual-view studies with complete Findings & Impression sections	Clinton K J	ClickUp
Text Cleaning	Applied RadGraph-based report filtering for high-quality annotations	Clinton K J	ClickUp
Fusion Module Design	Designed dual-view gated fusion architecture using CLIP, BiomedCLIP, and cross-attention	Clinton K J	ClickUp
Fine-Tuning	Fine-tuned Reasoning stage with prompt-based supervision	Clinton K J	ClickUp
Report Generation	Implemented Reporting module for full structured output	Clinton K J	ClickUp
Doc Review & QA	Reviewed report content and ensured technical consistency	Clinton K J	ClickUp
Image Preprocessing	Converted DICOM to JPEG, resized and normalized image pairs	Rahul Ramesh O	ClickUp
Prompt Design	Tuned instruction prompts for clinical reasoning	Rahul Ramesh O	ClickUp
Model Debugging	Assisted in analyzing training failures and resolving convergence issues	Rahul Ramesh O	ClickUp
Visualization Support	Helped prepare diagrams and prompt formatting samples	Rahul Ramesh O	ClickUp
Report Editing	Contributed to final report refinement and technical proofreading	Rahul Ramesh O	ClickUp
Documentation Co-author	Wrote the majority of the final report, including methodology and literature review	Rahul Ramesh O	ClickUp
Encoder Development	Developed encoder modules for frontal and lateral views	Farnaz Bigdeli	ClickUp

Report Writing	Reviewed report drafts and suggested formatting and text edits	Farnaz Bigdeli	ClickUp
Figure Preparation	Created key Figures and visual summaries for final report	Farnaz Bigdeli	ClickUp
Lit. Review Research	Conducted detailed review of prior models (PromptMRG, PMC-VQA, CheXFusion, etc.)	Farnaz Bigdeli	ClickUp
Architecture Diagram Design	Sketched and refined fusion/reasoning diagrams for visual documentation	Farnaz Bigdeli	ClickUp
Evaluation Summary Writing	Wrote interpretation and discussion for evaluation metrics (Table 1–3)	Farnaz Bigdeli	ClickUp
Feature Concatenation	Implemented concatenation of fused features from dual views	Kyatham Hemanth	ClickUp
Model Evaluation	Evaluated outputs using BLEU, ROUGE, METEOR, RadGraph F1, and CheXbert	Kyatham Hemanth	ClickUp
Results Compilation	Compiled qualitative examples and tabulated quantitative results	Kyatham Hemanth	ClickUp
Data Loader Optimization	Improved training efficiency by optimizing dual-view data loading pipeline	Kyatham Hemanth	ClickUp
Output Comparison Analysis	Analyzed and compared generated vs. ground-truth reports (highlighting alignments in Table 3)	Kyatham Hemanth	ClickUp
Documentation Co-author	Assisted in final proofreading and formatting of the report	Kyatham Hemanth	ClickUp

10. Appendix:

10.1 Hardware Configurations

- GPU model(s) used: NVIDIA A100
- Amount of GPU memory: 40 GB
- Number of GPUs used: 1
- CPU: 32 total cores
- RAM: 117 GB
- Root Disk: 60 GB
- Platform: Jetstream2

10.2 Platform/Tools Used

Programming Language

- Python

Frameworks / Libraries

- PyTorch – for deep learning model training
- Transformers – from Hugging Face, for loading and fine-tuning LLMs
- OpenCLIP – likely used as the vision tower for encoding images
- HuggingFace Datasets – for loading and processing datasets
- PEFT – for parameter-efficient fine-tuning
- BitsAndBytes – 8-bit optimizer for memory-efficient training
- WandB – for experiment tracking and logging
- Albumentations / Timm / torchvision – (likely) used for image augmentations and model utilities
- SentencePiece – for tokenization (if used in tokenizer pre-processing)

10.3. Data Description

Dataset Overview: MIMIC-CXR v2.0.0

MIMIC-CXR is a large-scale, publicly available dataset of chest radiographs released by MIT and PhysioNet. It includes:

- 377,110 chest X-ray images in DICOM format
- 227,827 free-text radiology reports
- From 65,379 unique patients
- Each study may contain one or more images (typically one frontal and optionally one lateral view)

The dataset is fully de-identified and comes with metadata like StudyInstanceUID, PatientID, ViewPosition, StudyDate etc.

Note: Access to the MIMIC-CXR dataset requires completion of a data use agreement (DUA) and credentialing process, including completion of human subjects research training. As a result, this dataset cannot be redistributed or made publicly available outside the terms of the original license and PhysioNet access process.

Data Preprocessing Steps

1. DICOM to JPEG Conversion

- All raw .dcm images were converted to .jpeg format using pydicom and OpenCV.
- The pixel values were processed with proper handling to ensure grayscale integrity.
- This conversion enabled compatibility with deep learning pipelines.

2. Image Cleaning and Resizing

- Black border removal was performed using pixel intensity thresholding to crop out uninformative edges.
- Images were resized to 224×224px while preserving aspect ratio to avoid distortion.
- Each image was then centered and padded to create uniform square dimensions.

3. Dual-View Study Selection

- For each study, StudyInstanceUID and ViewPosition were used to identify:
    - One frontal view
    - One lateral view
- Only studies with exactly one frontal and one lateral image were retained.

4. Study Filtering and Oversampling

Study Filtering:

The study includes a report-cleaning step using RadGraph, which filters out irrelevant or low-quality sentences based on the absence of clinically meaningful medical terms and entity relationships.

Oversampling:

- 76,323 clean dual-view studies for the reasoning stage

This set was oversampled 3×, specifically by generating varied prompts for the same images and reports, focusing on abnormalities.

- 92,215 image-report pairs for the reporting stage, also derived from oversampling.

5. Train/Test Splitting

- Dataset was split into train, validation, and test subsets with an approximate 95/5 ratio.
- Splitting was using study ids, ensuring no data leakage across splits.

10.4 Model Architecture and Hyperparameters

Vision Encoders Used

- CLIP Vision Tower: openai/clip-vit-large-patch14-336
- BiomedCLIP Vision Tower: Likely BiomedCLIP-PubMedBERT_256-vit_base_patch16_224
- These towers process frontal and lateral chest X-ray images (224×224) independently before fusion.
- Refer to Fig. 1: Dual-View Fusion (Architectural Contribution)

Language Model Used

- Base Model: Phi-2 (by Microsoft)
- Multimodal Model: LLaVA-Phi, a modified LLaVA architecture
- Trained progressively through:
  - Stage 1: Recognition
  - Stage 2: Reasoning
  - Stage 3: Reporting
  - Each stage fine-tunes the previous stage's model on new objectives. Refer to Fig. 2: Model Pipeline (Stage 1–3)

Fusion Mechanism Details

- Dual-View Fusion involves:
  - Feature extraction from both views using CLIP and BiomedCLIP
  - Medical Adapter Projection
  - Weighted Feature Addition
  - Cross-View Attention: frontal as query, lateral as key
  - Gated Fusion followed by a Multimodal Projection Layer
  - Outputs fed to LLM

See Fig. 3 for full fusion pipeline

Important Hyperparameters (All Stages)

Stage	Learning Rate	Batch Size*	Optimizer	Scheduler	Epochs	Weight Decay	Warmup Ratio	Max Grad Norm	Mixed Precision
Recognition	1e-5	4 (×8 = 32)	adamw_bnb_8bit	cosine	8	0.0	0.1	1.0	bfloat16 + TF32
Reasoning	5e-5	2 (×8 = 16)	same	cosine	2	0.0	0.1	1.0	bfloat16 + TF32

Reporting	2e-5	2 (×8 = 16)	same	cosine	3	0.0	0.03	1.0	bfloat16 + TF32

*Effective batch size = batch × gradient accumulation

10.5 Training Procedure

Epochs & Duration

Stage	Epochs	Approx. Training Time
Recognition	8	10 hours
Reasoning	2	26 hours
Reporting	3	40 hours

Loss Function

- Language modeling loss (auto-regressive next-token prediction)
- No contrastive or CLIP-style loss is used

Prompt Handling Differences

Stage	Prompt Version Used	Prompt Type	Pretrained?	Backbone Frozen
Recognition	plain	No user prompt used	Yes	Frozen
Reasoning	v0	Prompt-instructed questions	Finetuned	Frozen
Reporting	v0	Instruction: “Write a report”	Finetuned	Fully Unfrozen

Recognition JSON Sample (No prompt used during training)

```
{
 "frontal": "59502376/0.jpeg",
 "lateral": "59502376/1.jpeg",
 "conversations": [
   {
    "from": "human",
    "value": "<image> Prompt here (ignored in training)"
   },
   {
    "from": "gpt",
    "value": "Findings: PULMONARY VASCULAR REMAINS ENGORGED ..."
   }
 ]
}
```

Reasoning JSON Sample (Prompt-instructed differential diagnosis)

```
{
 "frontal": "58231510/0.jpeg",
 "lateral": "58231510/1.jpeg",
 "conversations": [
   {
    "from": "human",
    "value": "GIVEN THESE FINDINGS... <image> WHAT DIFFERENTIAL..."
   },
   {
    "from": "gpt",
    "value": "NO FOCAL CONSOLIDATION CONCERNING FOR PNEUMONIA."
   }
 ]
}
```

Reporting JSON Sample (Prompt-instructed report generation)

```
{
 "frontal": "56046385/0.jpeg",
 "lateral": "56046385/1.jpeg",
 "conversations": [
    {
     "from": "human",
     "value": "<image> Evaluate these CXRs and document your observations..."
    },
    {
     "from": "gpt",
     "value": "FINDINGS: OVERALL LUNG VOLUMES ARE LOW. ... IMPRESSION: ..."
    }
  ]
}
```

Data Oversampling Strategy

To improve learning from abnormal cases, **oversampling** was applied in later stages:

Stage	Oversampling Factor	Strategy Used	Final Size
Reasoning	×3 (abnormal only)	Repeated each abnormal case 3 times with different prompts	76,323 studies
Reporting	×3+ (abnormal only)	Same approach, for enhanced report variety	92,215 studies

This boosted diversity without changing ground truth, improving generalization in abnormal-case reasoning and generation.

10.6 Evaluation Metrics

General Language Quality Metrics

a. **BLEU (n-gram precision):**
   Formula: $\text{BLEU} = \text{BP} \times \exp(\Sigma (w_n \times \log p_n))$
   Where:
   ○ BP = brevity penalty
   ○ w_n = weight for each n-gram level
   ○ p_n = precision for n-gram matches

b. **ROUGE-L (Longest Common Subsequence Recall):**
   Evaluates the longest common subsequence between candidate and reference texts to assess fluency and structural similarity.

c. **METEOR (accounts for synonymy and word order):**
   Uses stemming, synonyms, and paraphrasing to align predicted and reference sentences beyond surface-level similarity.

Clinical Evaluation Metric

1. RadGraph F1 Score

The RadGraph dataset, released by the National Institutes of Health (NIH), is a richly annotated corpus of 2,400 English radiology reports sourced from the MIMIC-CXR and ChestXray14 datasets. It is designed to support the development and evaluation of clinical NLP and vision-language models. RadGraph provides detailed entity and relation annotations using a unified schema.

Each report is labeled with:

- Medical entities – e.g., *observations* and *anatomical locations*
- The relations between these entities

This structure captures clinical knowledge in graph format, enabling fine-grained evaluation for:

- Information extraction
- Radiology report generation
- Question answering in medical imaging

Dataset Statistics:

- Over 40,000 annotated entities
- Over 30,000 annotated relations
- Covers 14 entity types and 10 relation types
- Annotations performed by trained experts

RadGraph is commonly evaluated using the RadGraph F1 metric, which reflects how accurately models replicate both the clinical findings and their relationships.

RadGraph F1 Metric Working Details.

Entity Matching:

A predicted entity is considered correct if:

- The text span matches exactly
- The entity type matches
  (e.g., Observation, Anatomical Site)

Relation Matching:

A predicted relation is considered correct if:

- Both head and tail entities are correct
- The relation type matches
  (e.g., located_at)

Score Calculation:

- Precision = Correct predictions / Total predictions
- Recall = Correct predictions / Total ground truth
- F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

Final RadGraph F1 Score:

- Compute Entity F1 and Relation F1
- Take the average → Final RadGraph F1 (Range: 0 to 1)

2. CheXbert

CheXbert is an advanced radiology labeler built upon the CheXpert model, developed by Stanford University. It is designed to convert free-text chest X-ray radiology reports into structured binary labels across 13 common observations, including conditions like cardiomegaly, pneumonia, pleural effusion, and more.

Purpose:

CheXbert facilitates:

- Weak supervision for training radiology models.
- Automated evaluation of generated reports by comparing predicted labels to ground truth labels.

CheXbert Working Details

Label Extraction:

a. CheXbert uses BERT (Bidirectional Encoder Representations from Transformers), fine-tuned to classify radiology reports.
b. It extracts 13 predefined pathologies as binary labels (positive/negative) from the report text.
c. Also includes a "No Finding" label.

Inference Pipeline:

a. A report is input into the BERT-based classifier.
b. The model outputs 0 or 1 for each of the 13 labels.
c. For generated reports, these predicted labels are compared against labels extracted from the ground truth reports.

Evaluation Metric:

a. Label Accuracy Score: Measures how many labels from the predicted report match the ground truth.
i. Per-report Accuracy = (Number of correctly matched labels) / (Total number of labels)
ii. Final score is averaged across all reports.

Minutes of the Meeting (MoM)

Date: 8th July 2025
Time: 9:30 AM – 10:30 AM
Venue: Virtual Meeting
Attendees:
Great Learning Approval Committee (2 Members including Professor)

Team-15 Members:
Clinton K J
Rahul Ramesh O
Hemanth Kyatham
Farnaz

Discussion Summary

The meeting was conducted to review and provide feedback on the final project report submitted by Team-15. The committee shared valuable insights and highlighted necessary improvements to ensure the report aligns with institutional expectations and academic standards.

Key Feedback and Action Items

1. Originality Section Placement:
The originality declaration must be positioned at the beginning of the report for clarity and emphasis.

2. Template Compliance & Word Limit:
The team was advised to strictly follow the final report template format and adhere to the specified word limits, excluding tables, captions, appendix, bibliography, and individual contributions.

3. Member Contribution Section:
Details of individual contributions by team members should be moved to the very end of the report, not included in the main body.

4. Result Presentation:
All results should be presented in clearly formatted tables to enhance readability and maintain professionalism.

5. Language & Explanation Quality:
The committee recommended using more formal and refined language throughout the report. Bullet points should be converted into full explanatory sentences for a polished academic tone.

6. Dual View Architecture Explanation:
A detailed explanation of the dual-view architecture should be included, describing its purpose and how it enhances diagnostic performance by fusing features from both frontal and lateral chest X-rays.

7. Architecture Diagram Redesign:

The architecture diagram must be redrawn using smaller boxes to clearly indicate the team’s contributions to the base architecture. The design should aim for clarity, compactness, and effective visual representation.

8. Output Representation:
Model outputs must be shown in a clean and reader-friendly format. Screenshots or representations involving black terminal screens should be avoided.

9. Bibliography Inclusion:
A properly formatted bibliography section must be added at the end of the report to list all cited sources and references.