Assignment # 01

Advanced Artificial Intelligence

Name: Muhammad Talha

Roll No: 25K-7605

Program: MS(AI)

This document analyzes influential papers from 2023–2025 on automated chest X-ray report generation using vision–language models (VLMs). For each paper we summarize the approach, list pros and cons, and highlight gaps in the authors' problem statements. Using those insights we propose a formal research problem that directly addresses the recurring failure modes.

Papers analysed (shortlist)

Flamingo-CXR — clinical evaluation of a Flamingo‑based VLM fine-tuned for CXR report generation.

Knowledge‑enhanced Auto Diagnosis (KAD) — KAD exploits a medical knowledge graph during VLM pretraining/fine-tuning.

Longitudinal‑MIMIC / longitudinal prefill work — methods that use patient visit history (previous CXRs/reports) to prefill or condition report generation.

"Longitudinal data and a semantic similarity reward" (RL + longitudinal) — reward-based fine-tuning that simulates radiologist workflow.

2024 VLM survey / review — summarizes evaluation gaps, datasets and model families.

Figure 1.1 : Representing the core strengths and weaknesses of the papers

Paper / Year	**Approach**	**Strengths**	**Weaknesses**
ConVIRT (2020)	Contrastive VLP	Label-efficient, good transfer to new tasks	Report noise, poor localization performance
MoCo-CXR (2021)	Self-Supervised (MoCo)	Improves transfer in low-label regimes	Augmentations may erase subtle findings
GLoRIA (2021)	Global+Local Alignment	Better retrieval & interpretability, local alignment helps explainability	Attention maps ≠ true localization, weaker in edge cases
CheXzero (2022)	Zero-shot (CLIP-style)	Enables zero-shot diagnosis without retraining	Sensitive to prompt wording, weak on rare pathologies
MedCLIP (2022)	Semantic Matching VLP	Handles false negatives, supports unpaired data	Complex NLP pipeline needed, preprocessing heavy
REFERS (2022)	Cross-Supervised VLP	Strong results under label scarcity, effective representation learning	High compute cost, fragile to report text extraction

KAD (2023)	Knowledge-Enhanced VLP	Integrates medical knowledge, strong zero/few-shot	Depends on KB coverage, complex pipeline design
MedKLIP (2023)	Knowledge-Grounded VLP	Entity grounding + patch-level alignment improves interpretability	Hard to reproduce, complex engineering overhead
Comp. Study (2023)	Comparative Evaluation	Benchmarks multiple SSL/VLP approaches, clarifies strengths/weaknesses	Computationally heavy, results dataset-dependent
MIMIC-CXR Dataset	Dataset	Large paired image-report dataset, widely used benchmark	Single-center bias, noisy labels, limited global diversity

Proposed research problem

Title:

Knowledge-grounded Longitudinal Vision–Language Models for Clinically Faithful Chest X-Ray Report Generation.

Abstract:

Automated chest radiograph report generation holds promise to reduce clinician workload but is limited by hallucination, lack of clinical grounding, and poor generalization across institutions. We propose a hybrid approach that

(1) augments a vision–language foundation model with an explicit medical knowledge encoder,

(2) conditions generation on longitudinal patient history, and

(3) uses a clinically-informed reward during fine-tuning.

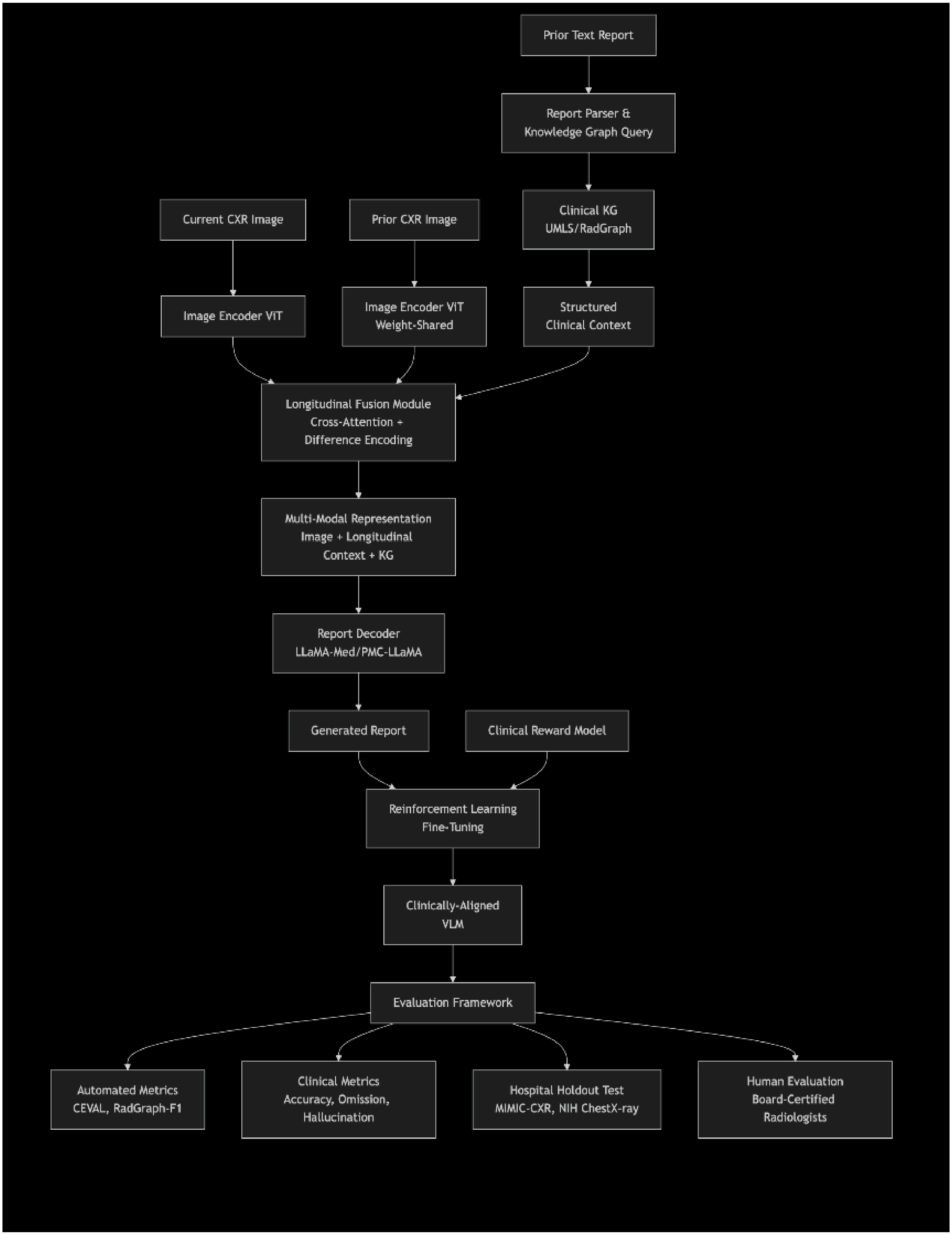
We evaluate the system on MIMIC-CXR (including Longitudinal-MIMIC splits), external hospital datasets, and via a blinded radiologist panel using a pre-registered evaluation protocol. We hypothesize that the combined interventions significantly reduce clinically significant errors while maintaining linguistic quality.

Formal Problem Statement: Given a chest X-ray image I_t and optional prior exams/reports {I_{t-1}, R_{t-1}, …}, learn a function G that generates a radiology report R^ that: (a) accurately reflects the current imaging findings; (b) correctly encodes temporal progression relative to priors; (c) aligns with structured clinical concepts in a medical knowledge graph; (d) minimizes clinically significant errors under external test distributions.

Objectives:

1. Design a VLM architecture that fuses image, longitudinal context, and KG-derived embeddings.
2. Define a clinically‑informed reward and apply it for fine-tuning without destabilizing generation.
3. Build and release evaluation splits (train/val/test + external hospital holdouts) with pre-registered clinical metrics and human evaluation protocol.
4. Quantify trade-offs between model size, compute cost, and clinical gains (a small‑model baseline vs a foundation VLM).

contd..


Prior Text Report
Report Parser & Knowledge Graph Query
Current CXR Image
Prior CXR Image
Clinical KG UMLS/RadGraph
Image Encoder ViT
Image Encoder ViT Weight-Shared
Structured Clinical Context
Longitudinal Fusion Module Cross-Attention + Difference Encoding
Multi-Modal Representation Image + Longitudinal Context + KG
Report Decoder LLaMA-Med/PMC-LLaMA
Generated Report
Clinical Reward Model
Reinforcement Learning Fine-Tuning
Clinically-Aligned VLM
Evaluation Framework
Automated Metrics CEVAL, RadGraph-F1
Clinical Metrics Accuracy, Omission, Hallucination
Hospital Holdout Test MIMIC-CXR, NIH ChestX-ray
Human Evaluation Board-Certified Radiologists

Research Questions (RQs):

- RQ1: Does knowledge grounding reduce clinically significant hallucinations compared to a vanilla VLM fine-tune?
- RQ2: Does longitudinal conditioning improve the accuracy of change descriptions (progression/regression) over single-exam models?
- RQ3: How does RL fine-tuning with a clinical reward compare to supervised fine-tuning with weighted loss terms?
- RQ4: What is the external generalization performance across hospital systems, and which components most improve robustness?

Hypotheses:

- H1: Knowledge grounding reduces the rate of clinically critical hallucinations by >X% (pre-specified threshold) vs baseline.
- H2: Longitudinal conditioning improves change-detection F1 by >Y points on longitudinal test splits.
- H3: Combined KG + longitudinal + clinical reward yields better clinical accuracy than any single intervention.