**Contrastive Pre-Training for Chest-X-ray Report Generation**

**Outline**

|  |  |
| --- | --- |
| **Rubric Section** | **Planned Content** |
| **INTRO / Goals** | Clinical motivation, problem statement, project objectives |
| **HYPOTHESES & METHOD** | Formal H₀/H₁, block diagram, algorithm math |
| **RESEARCH (Related Work)** | Prior datasets, contrastive & generative models, evaluation studies |
| **APPLICATION / Method & Solution** | Data pipeline, model architecture, metrics, hyper-parameter search |
| **WHAT IS LEARNED? / Conclusions** | Expected findings, clinical impact, future work |

**Problem, Candidate Solution, and Research Hypothesis:**

Chest radiograph interpretation is the most common imaging task worldwide, yet radiologist shortages delay care. Automated image-to-text systems can help, but current models rely heavily on supervised feature learning and often miss subtle findings on unseen institutions

**Candidate Solution:**

I am proposing to pre-train image encoders with self-supervised momentum-contrast (MoCo v3) on millions of unlabeled chest X-Rays, then fine-tune a vision-language model (Swin-T encoder + T5 decoder) on paired image-report data.

* H0 (null hypothesis): Contrastive pre-training yields no improvement over a purely supervised baseline
* H1 (alternative 1): Contrastive pre-training raises (a) BLEU-4, (b) ROUGE-L, and (c) *Clinical BERTScore* by ≥ 10% on the held-out MIMIC-CXR test set
* H2 (alternative 2): The model’s label-wise F1 for CheXpert observations improves by ≥ 5 percentage points (pp), which means it’s clinically meaningful.

**Applied ML Method & Dataset**

**Data:**

* **Pre-training**: Unlabeled images from **CheXpert** (224 k films) and **MIMIC-CXR** (377 k films)
* **Fine-tuning/ eval**: Paired image-report subsets of MIMIC-CSR and CheXpert validation set

**Architecture:**

1. **Image encoder**: Swin-Transformer-Tiny initialized with MoCo v3 contrastive weights
2. **Text decoder:** T5-Small initialized on general C4; cross-attention bridges vision tokens
3. **Training Strategy:**
   1. Stage 1: Momentum-contrast with queue = 65k, τ = 0.07
   2. Stage 2: Image-to-report fine-tune (mixed teacher forcing + label smoothing),
4. **Metrics**: BLEU-4, ROUGE-L, Clinical BERTScore, CheXpert-label F1, and radiologist error taxonomy survey.

**References:**

1. Irvin J. *et al.* “CheXpert: A Large Chest Radiograph Dataset …” arXiv:1901.07031,2019.
2. Johnson A. E. W. *et al.* “MIMIC-CXR …” *Sci. Data* 6, 2019.
3. Chen X. *et al.* “An Empirical Study of Training Self-Supervised Vision Transformers (MoCo v3).” ICCV 2021.
4. Zhang Y. *et al.* “MedCLIP: Contrastive Learning from Unpaired Medical Images and Reports.” EMNLP 2022.
5. RSNA. “Domain-specific Multimodal Generative AI Model for Chest Radiograph Report Generation.” *Radiology* 2025.
6. Tiu E. *et al.* “Evaluating Progress in Automatic Chest X-ray Report Generation.” *Patterns* 2023.
7. Cerri R. P. *et al.* “Vision-Language Models for Automated Chest X-ray Interpretation.” arXiv:2501.12356, 2025.
8. Yuan G. *et al.* “Clinical BERTScore: An Improved Measure …” ClinicalNLP 2023.
9. Stanford ML Group. “CheXpert Competition Website.” Accessed 2025.
10. Medium post “From MoCo v1 to v3: Dynamic Dictionary for SSL,” 2024 (background tutorial).