SRM Institute of Science and Technology
(Deemed to be University u/s 3 of UGC Act, 1956)
College of Engineering and Technology
School of Electrical and Electronics
Department of Electronics and Communication Engineering
21ECP401L-Major Project
AY 2025-2026

PROJECT PROPOSAL FORM

Project Title: Hierarchical Anatomical Query Transformer (HAQT) for Automated Chest X-Ray Report Generation

Supervisor : Dr. Damodar Panigrahy

Assistant Professor
Department of Electronics & Communication

Team Members : 1. S. Nikhil (Reg No: - RA2211004010362)
2. Dadhania Omkumar (Reg No: - RA2211004010392)
3. Rahul Saha (Reg No: - RA2211004010318)

Background/Literature Review:

Automated radiology report generation from medical images has emerged as a critical application of deep learning in healthcare. Current approaches face challenges in generating clinically accurate reports that properly describe findings across different anatomical regions.

Existing methods like R2Gen (Chen et al., EMNLP 2020) use memory-driven transformers but lack explicit anatomical awareness. METransformer (Wang et al., CVPR 2023) introduced multiple expert tokens but without spatial priors. RGRG (Tanida et al., CVPR 2023) uses region-guided generation but requires external object detectors. Recent vision-language models like BLIP-2 use Q-Former architectures for image-to-text generation but are not optimized for medical imaging where anatomical structure is critical.

Our proposed HAQT-ARR (Hierarchical Anatomical Query Tokens with Adaptive Region Routing) addresses these limitations by introducing learnable anatomical query tokens with 2D Gaussian spatial priors, enabling the model to focus on clinically relevant regions without requiring explicit segmentation masks at inference time.

References:

1. Wang, Z., Liu, L., Wang, L., & Zhou, L. (2023). "METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens." IEEE/CVCVPR, pp.11558-11567.

2. Tanida, T., Muller, P., Kaissis, G., & Rueckert, D. (2023). "Interactive and Explainable

Region-guided Radiology Report Generation." IEEE/CVF CVPR 2023.

3. Wang, L., et al. (2023). "R2GenGPT: Radiology Report Generation with Frozen LLMs." Meta-Radiology, 1(3):100033. arXiv:2309.09812.

4. Authors (2025). "ChestX-Transcribe: A Multimodal Transformer for Automated Radiology Report Generation." Frontiers in Digital Health, 2025.

5. Authors (2024). "Energy-Based Controllable Radiology Report Generation with Medical Knowledge." MICCAI 2024.

6. Liu, Z., et al. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." IEEE/CVF ICCV 2021.

7. Yuan, H., et al. (2022). "BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model." BioNLP Workshop, ACL 2022.

8. Johnson, A. E., et al. (2019). "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." Scientific Data.

9. Li, J., et al. (2023). "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." ICML 2023.

10. Chen, Z., et al. (2020). "Generating Radiology Reports via Memory-driven Transformer." EMNLP 2020.

Note: Minimum 5 recent references must be given

Related course completed in previous semester with subject code and name: Nil

Objective:

To develop an end-to-end deep learning system for automated generation of clinical radiology reports from chest X-ray images using a novel HAQT-ARR (Hierarchical Anatomical Query Tokens with Adaptive Region Routing) architecture that :
1. Extracts hierarchical visual features using Swin Transformer encoder
2. Projects visual features to language space using anatomical-aware query tokens with learnable spatial priors
3. Generates clinically accurate reports using BioBART decoder pre-trained on biomedical text
4. Provides interpretable attention visualization showing which anatomical regions influenced the generated report

Requirements:

1) NVIDIA RTX 4060 Laptop GPU (8GB VRAM)
2) 16GB System RAM, 512GB SSD Storage
3) Python 3.10+ with PyTorch 2.1+ and CUDA 12.1
4) Transformers (HuggingFace) 4.36+, timm 0.9+ (Pre-trained Vision Model Library)
5) FastAPI for REST API Backend
6) React 18 with TypeScript for Frontend
7) Node.js 18+ for Frontend Build
8) MIMIC-CXR Dataset (30,633 chest X-ray images)

9) Languages: Python, TypeScript, CUDA
10) Development: VS Code, Jupyter Notebooks

Technical Requirements:

Vision Encoder: Swin Transformer Base (87M parameters, pretrained on ImageNet)
Projection Layer: HAQT-ARR Novel (36 query tokens: 8 global + 7x4 anatomical regions)
Text Decoder: BioBART (139M parameters, pretrained on PubMed biomedical literature)
Training: Mixed Precision FP16, 50 epochs, effective batch size 32
API Server: FastAPI + Uvicorn (RESTful endpoints for inference)
Frontend: React + Vite + Tailwind CSS (Interactive dashboard with attention visualization)

Engineering standards and realistic constraints:

Area	Codes & Standards / Realistic Constraints
Economic	Targeted maximum expense of Rs 15,000/- (GPU cloud compute if needed). Primary development on personal laptop with RTX 4060 GPU.
Environmental	Project involves computational workload; GPU training estimated at ~30 hours total. Energy-efficient mixed precision (FP16) training employed to reduce power consumption.
Social	Project aims to assist radiologists by reducing report generation time from minutes to seconds, improving healthcare accessibility in resource-limited settings with radiologist shortages.
Ethical	Uses de-identified MIMIC-CXR dataset (publicly available). AI-generated reports clearly marked as requiring physician review. No autonomous clinical decisions made by the system.
Health and Safety	Software-only project with no physical health risks. Generated reports explicitly marked as AI-assisted, not diagnostic. Not intended for clinical use without radiologist validation.
Manufacturability	Fully reproducible: Complete source code on GitHub, trained model checkpoints, comprehensive documentation, and Jupyter notebooks provided. Deployable via Docker containers.
Sustainability	Cloud-deployable solution reducing need for physical infrastructure. Model weights shareable for research reuse. Open-source codebase for community contributions and improvements.

Realistic Constraints:

1.) Computational Constraints: Training limited to RTX 4060 (8GB VRAM), using gradient accumulation and mixed precision. Dataset reduced from 370K to 30K curated subset.
2.) Clinical Validation: Generated reports require validation by certified radiologists before clinical deployment. Current evaluation uses NLG metrics (BLEU, ROUGE) as proxies.

3.) Model Interpretability: HAQT-ARR provides attention visualization but does not guarantee clinical explainability required for regulatory approval (FDA).

Deliverables:

1) 1. Trained XR2Text Model with HAQT-ARR projection layer (~251M parameters, checkpoint files)
2) 2. Backend API Server (FastAPI) with endpoints for report generation, batch processing, attention visualization, and feedback collection
3) 3. Frontend Dashboard (React) with X-ray upload, real-time report generation, anatomical attention visualization, and report history
   4. Jupyter Notebooks for data exploration, model training, evaluation metrics, and ablation studies

Standards Referred/used:

1) DICOM - Digital Imaging and Communications in Medicine (Medical Image Format)
2) HL7 FHIR - Healthcare Interoperability Standard
3) IEEE 29148-2018 - Systems and Software Engineering Requirements

Abstract:

Automated radiology report generation from chest X-ray images is a critical task that can significantly reduce radiologist workload and improve diagnostic efficiency. This project presents XR2Text, a novel end-to-end transformer-based system featuring HAQT-ARR (Hierarchical Anatomical Query Tokens with Adaptive Region Routing) - a new projection mechanism that bridges vision and language models with explicit anatomical awareness.

Unlike existing approaches that use generic query tokens or require external organ detectors, HAQT-ARR introduces learnable 2D Gaussian spatial priors for seven anatomical regions (lungs, heart, mediastinum, spine, diaphragm, costophrenic angles) and dynamically routes attention based on image content. The system combines a Swin Transformer encoder for hierarchical visual feature extraction with a BioBART decoder pre-trained on biomedical literature for clinically accurate text generation. Additional novel contributions include: (1) Anatomical-aware curriculum learning progressing from normal to complex cases, (2) Custom loss functions for clinical entity detection and anatomical consistency, and (3) A clinical validation framework for automated accuracy assessment. Trained on 30,633 MIMIC-CXR images using an RTX 4060 GPU, the system achieves competitive performance on standard NLG metrics while providing interpretable attention visualizations.

Additional Requirements:

(Multidisciplinary tasks – Computational /IT involved, Biomedical Engineering, **)**

This project involves deep learning model development (Computer Science), web application development (Information Technology), medical image processing

(Biomedical Engineering), and statistical analysis for evaluation metrics (Data Science/Mathematics).

Other Department	Utilised for	Remarks
Basic Sciences	Linear Algebra, Probability	Transformer attention mechanisms
Mechanical Engineering	Nil	Nil
Instrumentation and Control Engineering	Nil	Nil
Electrical and Electronics Engineering	Signal Processing	Image preprocessing
Computational/IT	PyTorch, React, FastAPI	Deep Learning & Web Dev
Biomedical Engineering	Medical Imaging, DICOM	Clinical terminology
Purchase Section	Nil	Nil
Maintenance Department	Nil	Nil
Desktop publications	Project Report, Documentation	LaTeX/Word formatting

Design Project Summary

Project Title	Hierarchical Anatomical Query Transformer (HAQT) for Automated Chest X-Ray Report Generation
Objective of the Project	To develop an AI system that automatically generates clinical radiology reports from chest X-ray images using novel HAQT-ARR architecture with anatomical awareness and interpretable attention visualization.
Realistic constraints imposed	1.) Computational: Limited to RTX 4060 8GB GPU, using 30K image subset 2.) Clinical: Reports require radiologist validation before

	clinical use 3.) Regulatory: Not FDA-approved for clinical deployment 2.) Despite high cost it is more cost effective than systems using CDMA etc. 3.) Speed is high.
Standards to be referred/followed	1) DICOM - Medical Imaging Format HL7 FHIR - Healthcare Interoperability IEEE 29148 - Requirements Engineering HIPAA - Data Privacy Guidelines 2) IEEE 802.15.4- ZigBee 3) IS 12970- Semiconductor devices Integrated circuits
Multidisciplinary tasks involved	1) Computer Science for Deep Learning (PyTorch, Transformers) Information Technology for Web Development (React, FastAPI) Biomedical Engineering for Medical Imaging (DICOM, Clinical NLP) Data Science for Evaluation Metrics (BLEU, ROUGE, Statistical Tests) 2) Computational and IT field for Visual basic 6.0. 3) Desktop publication for report.

Supervisor's Signature