DAYANANDA SAGAR UNIVERSITY

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt,
Bengaluru-562112, Karnataka, India

SCHOOL OF ENGINEERING

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

A Project Report On

Enhancing Medical Diagnostics with Vision-Language Models

By

Shriyans Shriniwas Arkal - ENG21AM0117
Sri Bharath Sharma P - ENG22AM3005
Yudhajit Jana - ENG22AM3021

Under the supervision of
Dr. Vinutha N
Associate Professor
Computer Science & Engineering (AI & ML)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

SCHOOL OF ENGINEERING
DAYANANDA SAGAR UNIVERSITY

(2024 – 2025)

DAYANANDA SAGAR UNIVERSITY

SCHOOL OF ENGINEERING

Department of Computer Science & Engineering

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt, Bengaluru-562112, Karnataka, India

CERTIFICATE

This is to certify that the project entitled **"Enhancing Medical Diagnostics with Vision-Language Models"** is carried out by **Shriyans Shriniwas Arkal (ENG21AM0117), Sri Bharath Sharma P (ENG22AM3005), Yudhajit Jana (ENG22AM3021)**, bonafide students of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore, in partial fulfillment for the award of a degree in Bachelor of Technology in Computer Science and Engineering, during the year **2024 - 2025**.

Dr. Vinutha N	**Dr. Vinutha N**	**Dr. Jayavrinda Vrindavanam**
Associate Professor	Project Co-ordinator	Professor & Chairperson
Dept. of CSE (AI&ML)	Dept. of CSE (AI&ML)	Dept. of CSE (AI&ML)
School of Engineering	School of Engineering	School of Engineering
Dayananda Sagar University	Dayananda Sagar University	Dayananda Sagar University
Signature	Signature	Signature .....

Name of the Examiners:	Signature with date:
1.....	.....
2.....	.....
3.....	.....

DECLARATION

We, **Shriyans Shriniwas Arkal (ENG21AM0117), Sri Bharath Sharma P (ENG22AM3005), Yudhajit Jana (ENG22AM3021)**, are students of the eighth semester B.Tech in Computer Science and Engineering (AI & ML) at the School of Engineering, Dayananda Sagar University. We hereby declare that the Major Project titled **"Enhancing Medical Diagnostics with Vision-Language Models"** has been carried out by us and submitted in partial fulfillment for the award of a degree in **Bachelor of Technology** in **Computer Science and Engineering** during the academic year **2024–2025**.

Student: **Signature**

Name 1: Shriyans Shriniwas Arkal
USN: ENG21AM0117

Name 2: Sri Bharath Sharma P
USN: ENG22AM3005

Name 3: Yudhajit Jana
USN: ENG22AM3021

Place: Bangalore
Date:

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work. First, we take this opportunity to express our sincere gratitude to School of Engineering & Technology, Dayananda Sagar University for providing us with a great opportunity to pursue our Bachelor's degree in this institution.

We would like to thank **Dr. Udaya Kumar Reddy K R, Dean, School of Engineering & Technology, Dayananda Sagar University** for his constant encouragement and expert advice. It is a matter of immense pleasure to express our sincere thanks to **Dr. Jayavrinda Vrindavanam, Department Chairman, Computer Science and Engineering (Artificial Intelligence and Machine Learning), Dayananda Sagar University**, for providing right academic guidance that made our task possible.

We would like to thank our guide **Dr. Vinutha N, Associate Professor, Dept. of Computer Science and Engineering (Artificial Intelligence and Machine Learning) Dayananda Sagar University**, for sparing her valuable time to extend help in every step of our project work, which paved the way for smoothprogress and fruitful culmination of the project.

We would like to thank our **Project Coordinator Dr. Vinutha N** as well as all the staff members of **Computer Science and Engineering (Artificial Intelligence and Machine Learning)** for their support. We are also grateful to our family and friends who provided us with every requirement throughout the course. We would like to thank one and all who directly or indirectly helped us in the Project work

Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Generative AI, particularly in the form of multi-modal large language models (MLLMs) and vision-language models (VLMs), is reshaping the landscape of medical diagnostics and image interpretation. This work presents the development of a domain-adapted vision language model designed for radiological analysis, leveraging a unified transformer-based architecture that integrates a vision backbone and a text encoder. Visual and textual embeddings are jointly processed and fused through multimodal attention mechanisms to generate clinically relevant and contextually accurate outputs.

To address the limitations of general-purpose vision models in medical imaging tasks, the LLaMA 3.2-11B-Vision model was fine-tuned using Low-Rank Adaptation (LoRA) on the ROCOv2 (Radiology Objects in Context) dataset. Evaluation was conducted using DeepEval metrics configured for clinical relevance. The fine-tuned model achieved a G-Eval score of 0.75, a Hallucination Detection score of 0.71, and a Faithfulness score of 0.79, significantly outperforming the base model scores of 0.27, 1.00, and 0.69, respectively. Additionally, a BERTScore F1 of 0.85 indicates strong semantic similarity to expert-generated captions. These results demonstrate that the proposed system not only improves accuracy and reliability in radiological interpretation but also sets a foundation for integrating vision language models into real-time clinical workflows, offering healthcare professionals a powerful tool for enhanced diagnostic support.

Chapter 1

INTRODUCTION

Vision-language models (VLMs) are transforming clinical imaging by enabling computers to interpret medical images and communicate findings using the specialised language of radiology. A foundational advancement came with **MedFlamingo**, where a 9-billion parameter OpenFlamingo model, given only a few in-context examples, achieved near-expert performance in answering open-ended visual questions. Building on this, subsequent models streamlined vision-language integration: **XrayGPT** connected a frozen MedCLIP encoder to a Vicuna LLM with a single linear layer; **PubMedCLIP** focused on contrastive retrieval; and **BioViL-T** introduced temporal reasoning across sequential exams. More recently, **LLaVA-Med** demonstrated that instruction-tuning a 7B LLaMA model on GPT-4-generated dialogues can outperform fully supervised visual question answering (VQA) systems. LoRA-based adaptations have further expanded this efficiency-driven tuning approach to specialised domains like mammography.

- **Context:** Despite these advances, several challenges limit real-world deployment of medical VLMs. Many models rely on loosely coupled vision and language components or use smaller backbones, reducing cross-modal reasoning depth. Most are limited to single imaging modalities and static images, hindering generalisation across modalities or over time. Full fine-tuning of large models remains computationally prohibitive for most healthcare setups, while few-shot methods depend on carefully designed exemplars rarely available in routine workflows.
- **Objective:** This work aims to bridge these gaps by adapting the latest state-of-the-art **LLaMA-3.2-11B-Vision-Instruct** architecture for the field of radiology. Leveraging **Low-Rank Adaptation (LoRA)** within the highly efficient **Unsloth** training framework, we fine-tune the model for radiological tasks with minimal parameter updates. This enables high-performance medical inference without the need for resource-intensive retraining.

- **Scope:** Our study utilises the **ROCOv2** dataset, a rich and diverse corpus comprising approximately 1978 medical images along with expert-generated captions. The dataset includes imaging modalities such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and others. By updating less than 1% of the total model parameters, our LoRA-based fine-tuning approach reduces GPU memory consumption by approximately 60% in comparison to conventional full model retraining. This makes our system highly accessible to institutions equipped with only workstation-class GPUs.

- **Significance:** The resulting unified model architecture offers an end-to-end solution that combines the linguistic depth of a powerful large language model with a comprehensive visual interface. Clinicians can now interact with the system to pose open-ended diagnostic queries, solicit differential diagnoses, or even generate simulated clinical teaching cases across multiple imaging modalities—all without the burden of providing hand-crafted prompt exemplars. This paradigm shift enhances the usability and generalisation of VLMs in practical radiological settings.

- **Performance Evaluation:** The system's performance is rigorously evaluated across multiple standardised benchmarks. We employ a combination of **BERTScore-F1**, **DeepEval composite metrics**, **METEOR**, **ROUGE-L**, and **BLEU** to assess model output across four major tasks: image captioning, medical concept detection, and question answering on the **VQA-RAD** and **PathVQA** datasets. The results demonstrate a strong alignment between generated outputs and reference ground truths, while maintaining high semantic precision. Notably, the system operates effectively on a single GPU, making it suitable for deployment in typical clinical environments.

- **Reproducibility and Accessibility:**
  To promote adoption and further innovation, we make available a complete suite of resources. This includes:

  - The LoRA delta weights for parameter-efficient adaptation
  - Comprehensive preprocessing scripts for data alignment and formatting
  - A one-command training recipe through Unsloth that allows for seamless reproduction of results

  These resources significantly lower the entry barrier for healthcare researchers and practitioners aiming to develop or fine-tune their own domain-specific VLMs.

Chapter 2

PROBLEM DEFINITION AND OBJECTIVES

2.1 Problem Definition

Accessing precise and contextually relevant medical insights from complex radiological and pathological data poses a significant challenge for healthcare professionals. While existing vision-language models offer general-purpose capabilities, they often fail to perform effectively on domain-specific tasks due to a lack of fine-tuning for specialized medical datasets. This limitation leads to suboptimal diagnostic accuracy and restricts the practical application of AI in medical workflows. The absence of domain-adapted models hampers the reliability of AI-driven diagnostic support systems, slowing their adoption in clinical practice and limiting their potential to revolutionize healthcare.

2.2 Objectives

To address this gap, the LLaMA-3.2-11B-Vision-Instruct backbone is fine-tuned with Low-Rank Adaptation (LoRA) on the ROCOv2 corpus and related medical image-text datasets, updating less than 1% of its parameters to inject radiology-specific knowledge while maintaining computational efficiency. Model optimisation targets caption generation, concept detection, and visual QA, with performance assessed via BERTScore-F1, DeepEval, METEOR, ROUGE-L, and BLEU to ensure linguistic fidelity and semantic accuracy.

The resulting system supports open-ended clinical dialogue—producing differential diagnoses, rationale explanations, and simulated teaching cases—and operates on a single workstation-class GPU. LoRA weight deltas, preprocessing scripts, and a one-command Unsloth training recipe are released to enable transparent replication and straightforward extension to additional medical sub-domains.

Chapter 3

LITERATURE SURVEY

An Introduction to Vision-Language Modeling [4] by Florian Bordes et al. mentions that Vision Language Models (VLMs) represent an exciting frontier in AI, extending the capabilities of Large Language Models (LLMs) to the visual domain. These models have transformative potential, enabling applications ranging from advanced visual assistants to generative models that produce images from text prompts. However, the multimodal nature of vision and language poses unique challenges, as language operates in a discrete format while vision encompasses a higher-dimensional, continuous space. Vision-language model training employs diverse paradigms such as contrastive methods, masking strategies, generative components, and leveraging pre-trained backbones like LLMs. Key considerations include curating high-quality, diverse datasets, optimizing data pruning to align captions with images, and improving generative components for enhanced vision understanding. Grounding—mapping text with visual clues—and alignment with human expectations are central to vision-language model effectiveness. Despite the computational expense, with training often requiring large-scale GPUs and millions of image-text pairs, innovations such as using pre-trained language models and visual extractors offer cost-efficient solutions. Evaluating vision-language models remains complex, as many benchmarks rely heavily on language priors, raising concerns about genuine visual understanding. Extending vision-language models to video introduces further computational and data challenges, highlighting the ongoing need for robust methodologies and scalable solutions in this evolving field.

LLAMAFACTORY: Unified Efficient Fine-Tuning of 100+ Language Models [22] by Yaowei Zheng et al. states that LLAMAFACTORY is an advanced framework designed to streamline the fine-tuning of Large Language Models (LLMs) for downstream tasks with limited resources, integrating state-of-the-art efficient training techniques. Its modular design comprises a Model Loader for handling model initialization, quantization, and adapter attachment, a Data Worker for standardizing and preprocessing diverse datasets, and a Trainer for implementing fine-tuning approaches like pre-training, instruction tuning, and preference optimization. The LLAMABOARD interface, built on Gradio, enhances accessibility by enabling code-free customization of fine-tuning processes.

LLAMAFACTORY supports techniques for efficient optimization, such as LoRA, QLoRA, and PiSSA, alongside efficient computation methods like mixed precision training and flash attention. The framework supports over 100 LLMs, including LLaMA and Llama 2, and diverse datasets for tasks like language modeling and text generation. Empirical evaluations using datasets such as PubMed and CNN/DM showcase LLAMAFACTORY's efficiency in memory usage, throughput, and task-specific performance metrics like perplexity and ROUGE scores. With its user-friendly design and extensive capabilities, LLAMAFACTORY has democratized LLM customization, contributing to the growth of open-source LLM research. Future plans include support for multimodal models, advanced conversational fine-tuning techniques, and parallel training strategies, solidifying its role as a versatile and impactful tool in the field.

LoRA: Low-Rank Adaptation of Large Language Models [9] by Edward Hu et al. explores Low-Rank Adaptation (LoRA), a technique that addresses the challenges of adapting large language models (LLMs) to downstream tasks by overcoming the high computational and storage demands of full fine-tuning. Existing methods, such as adapter layers and prompt tuning, are discussed, highlighting their limitations, including inference latency introduced by adapter layers. LoRA emerges as a solution by injecting trainable rank decomposition matrices into each Transformer layer, which reduces the number of trainable parameters, lowers storage requirements, accelerates training, and eliminates additional inference latency. Empirical evaluations across models like RoBERTa, DeBERTa, GPT-2, and GPT-3 demonstrate that LoRA achieves performance on par with or better than full fine-tuning while being significantly more efficient. The review also examines the properties of low-rank updates, analyzing the relationship between the adaptation matrix (W) and pre-trained weights (W), and suggests that W amplifies underutilized features from pre-training, revealing insights into LLM adaptation mechanisms. Concluding with the advantages of LoRA for efficient and effective LLM fine-tuning, the review proposes future directions, such as combining LoRA with complementary techniques and deepening understanding of its adaptation mechanisms.

SWIFT: A Scalable lightWeight Infrastructure for Fine-Tuning [21] by Yuze Zhao et al. states that SWIFT is an open-source framework that addresses the challenges of training and deploying large language models (LLMs) and multi-modal large language models (MLLMs), offering a comprehensive solution spanning pre-training, fine-tuning, alignment, quantization, inference, evaluation, and deployment. Developed to streamline the resource-intensive process associated with large models, SWIFT builds upon advancements like Low-Rank Adaptation (LoRA), rsLoRA, and quantization techniques while providing a unified interface for efficient model customization and deployment. Supporting over 300 LLMs and 50 MLLMs, it integrates libraries such as PEFT and Optimum to enable cutting-edge fine-tuning and post-training operations like quantization and evaluation. SWIFT includes features like a unified training interface for both text and multi-modal models, extensive evaluation capabilities through the EvalScope framework, and user-friendly command-line and web interfaces.

Experimental validation highlights its effectiveness in lightweight tuning and agent training, demonstrating significant resource savings and improved model performance, such as reduced hallucinations and enhanced task accuracy. Future directions include expanding support for large-scale parallel training, advancing multi-modal research, and integrating with Retrieval Augmented Generation (RAG) systems. By democratizing the development of LLMs, SWIFT serves as a powerful and accessible tool for leveraging the potential of these transformative technologies.

ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset [16] by Rückert et al. introduces a comprehensive multimodal dataset designed to advance medical AI research. Building on the original ROCO dataset, ROCOv2 provides 79,789 radiological images paired with detailed textual captions and Unified Medical Language System (UMLS) concepts, enabling tasks such as concept detection, image-caption matching, and caption prediction. Unlike datasets such as MIMIC-CXR and Open-I, which focus predominantly on chest X-rays, ROCOv2 encompasses multiple imaging modalities, including CT, MRI, PET, ultrasound, and X-rays, across diverse anatomical regions. This broader scope makes ROCOv2 a versatile resource for training and evaluating multimodal models in the medical domain. To ensure high-quality annotations, Rückert et al. combine manual validation with automatic concept extraction using tools like MedCAT, linking captions to relevant UMLS concepts while filtering out noise and ambiguities. Additionally, the dataset introduces curated concepts for body regions and directionality specific to X-ray images, facilitating more detailed anatomical understanding. Challenges such as caption incompleteness, directional ambiguities, and modality imbalances reflect the inherent complexity of medical data curation, yet ROCOv2 demonstrates significant improvements over earlier datasets. By enabling tasks like multimodal retrieval, concept-based detection, and generative reporting, ROCOv2 sets a new benchmark for developing and evaluating deep learning models tailored to radiology. Future work may address remaining limitations, such as ambiguous anatomical labels, while expanding coverage of underrepresented modalities to enhance the dataset's robustness and utility.

Chapter 4

METHODOLOGY

4.1 Architecture of Proposed System

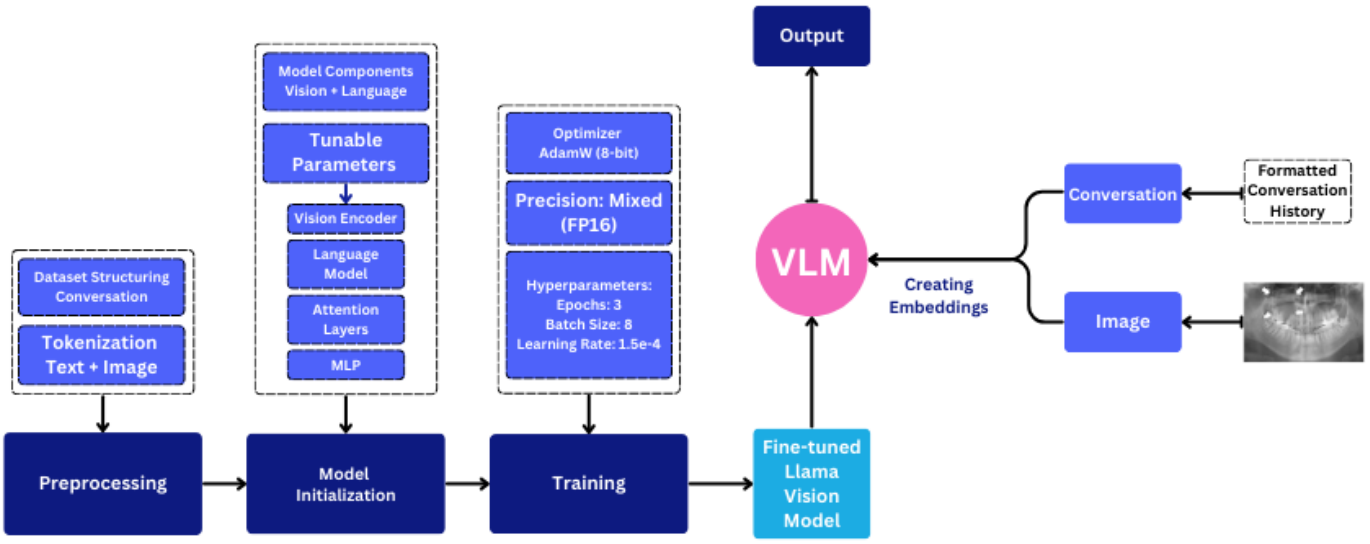


Figure 4.1: Project Architecture Overview

4.1.1 Data Preprocessing

Dataset Structure: The dataset used for this fine-tuning task is **ROCOv2**, a comprehensive multimodal dataset tailored for advancing medical AI research. It includes:

- **Images:** High-resolution radiological imagery across multiple modalities such as X-rays, CT scans, MRIs, PET scans, and ultrasounds. These images represent a wide array of anatomical regions, pathological findings, and normal observations, ensuring diverse coverage for robust model training and evaluation.
- **Captions:** Expert-provided textual descriptions paired with each image, detailing observations of medical conditions or normalities.

These captions are enriched with **Unified Medical Language System (UMLS)** concepts, enabling advanced tasks such as concept detection, image-caption matching, and caption prediction.

Table 4.1: Dataset Schema

Column Name	Type	Units	Description
`image`	image	width (px)	High-resolution medical image.
`image_id`	string	lengths	Unique identifier for the image.
`caption`	string	lengths	Textual description of the image.
`cui`	sequence	lengths	Concept Unique Identifier (CUI) *(Dropped).*

Instruction Preparation: To guide the Vision-Language Model (VLM) effectively during fine-tuning and inference, custom instructions are meticulously crafted to ensure the model behaves as an expert radiographer. These instructions are designed to provide the model with clear and precise guidance on how to interpret medical images and generate contextually accurate and clinically relevant captions. For instance, the instructions may include prompts to focus on specific anatomical regions, identify abnormalities, or describe findings using standardized medical terminology. By tailoring these instructions, the model is better equipped to understand the nuances of medical imaging and deliver outputs that align with the expectations of healthcare professionals. This step is critical to ensure that the model's outputs are both reliable and actionable in a clinical setting.

Pre-fine-tuning Inference: Before initiating the fine-tuning process, the pre-trained model's performance is rigorously evaluated on the target dataset. This evaluation serves as a benchmark to assess the model's initial ability to analyze medical images and generate captions. The process involves running the model on a subset of the dataset and comparing its outputs against ground truth annotations or expert-generated captions. Key metrics such as accuracy, precision, recall, BLEU scores, or clinical relevance are calculated to identify strengths and weaknesses in the model's current capabilities. This step provides valuable insights into specific areas that require improvement during fine-tuning, such as enhancing the model's ability to detect subtle abnormalities or improving its understanding of domain-specific terminology. The results of this evaluation guide the subsequent fine-tuning process, ensuring that training efforts are focused on addressing the model's limitations.

Training Preparation:

- **Dataset Transformation:** The dataset is carefully pre-processed and transformed into a format compatible with the model's input requirements. This involves resizing and normalizing medical images, tokenizing text annotations, and creating paired image-text datasets. Additionally, data augmentation techniques are applied to increase the diversity of the training data, such as flipping, rotating, or adjusting the contrast of images.

These transformations ensure that the model is exposed to a wide variety of scenarios, improving its robustness and generalization capabilities.

- **Training Configuration:** Training configurations are carefully set to optimize the fine-tuning process. Parameters such as batch size, learning rate, and optimization strategy are chosen based on the computational resources available and the complexity of the dataset. For instance, smaller batch sizes may be used to manage memory constraints when working with high-resolution medical images, while adaptive learning rates can help the model converge more effectively. Regularization techniques, such as dropout or weight decay, may also be employed to prevent overfitting. These configurations are critical to ensuring that the training process is both efficient and effective, enabling the model to achieve high performance without exceeding resource limitations.

Dataset Conversion: The dataset is reformatted into a multimodal conversational structure to align with the model's input expectations. Each example is structured as follows:

```
{
  "messages": [
    {
      "role": "user",
      "content": [
        { "type": "text", "text": "You are an expert radiographer.
        Describe accurately what you see in this image." },
        { "type": "image", "image": sample["image"] }
      ]
    },
    {
      "role": "assistant",
      "content": [
        { "type": "text", "text": sample["caption"] }
      ]
    }
  ]
}
```

This step ensures the model receives both visual and textual context in a structured, dialogue-based format.

4.1.2 Architecture Workflow

1. **Data Preprocessing**

   - *Loading Dataset:* The ROCO Radiology dataset is imported via Hugging Face, and the relevant split (e.g., training data) is selected. For this study, we utilized a sample of 1,978 entries from the approximately 60,000 available in the training split.
   - *Data Inspection:* The dataset structure is examined to understand its content and potential preprocessing requirements.
   - *Data Transformation:* Each data entry is converted into a multimodal conversation format, consisting of user instructions and corresponding image-text pairs.
   - *Data Formatting:* The dataset is structured into a format suitable for vision-language model fine-tuning, adhering to the template required by the Unsloth framework.

2. **Model Training**

   - *Model Initialization:*
     - The pre-trained Llama-3.2 Vision model and its tokenizer are loaded.
     - Gradient checkpointing is enabled for memory efficiency, and LoRA (Low-Rank Adaptation) adapters are applied for parameter-efficient fine-tuning.
   - *Training Setup:*
     - **Batch size:** 8 samples per device.
     - **Gradient Accumulation:** Steps set to 2 for memory and throughput balance.
     - **Learning Rate:** 1.5e-4 for stable optimization.
     - **Epochs:** 3 for comprehensive dataset coverage.
     - **Warmup Steps:** 10 for gradual model adaptation.
     - **Weight Decay:** 0.01 for regularization.
   - *LoRA Adapter Configuration:*
     - Vision-specific, language-specific, attention, and MLP modules are all fine-tuned.
     - Hyperparameters include LoRA rank (r) of 32, an alpha of 32, and a dropout of 0.1.
   - *Training Execution:*
     - The training process is conducted using Hugging Face's `TRL SFTTrainer` and the custom `UnslothVisionDataCollator`.
     - Real-time monitoring of GPU memory ensures efficient resource utilization on an NVIDIA H100 80GB SXM4 GPU.

3. **Model Inference and Sampling**

- *Input Preparation:*
  - Input consists of medical images and corresponding instructions (e.g., "Describe accurately what you see in this image, without patient details.").
  - Data is tokenized and structured for model inference.
- *Text Generation:*
  - The model generates outputs with fine-tuned parameters for creativity and accuracy, using a minimum probability (`min_p`) of 0.1 and temperature of 1.5.
  - A text streamer captures real-time outputs for evaluation.
- *Post-Processing:*
  - Outputs are optionally formatted to align with clinical standards or specific use cases.

4.2 Model Used

The foundation of this project is the **LLaMA 3.2-11B Vision-Instruct** model, a state-of-the-art multimodal architecture developed by Meta AI. This model is designed to process and understand both textual and visual inputs, making it suitable for tasks such as image captioning, visual question answering, and document analysis.

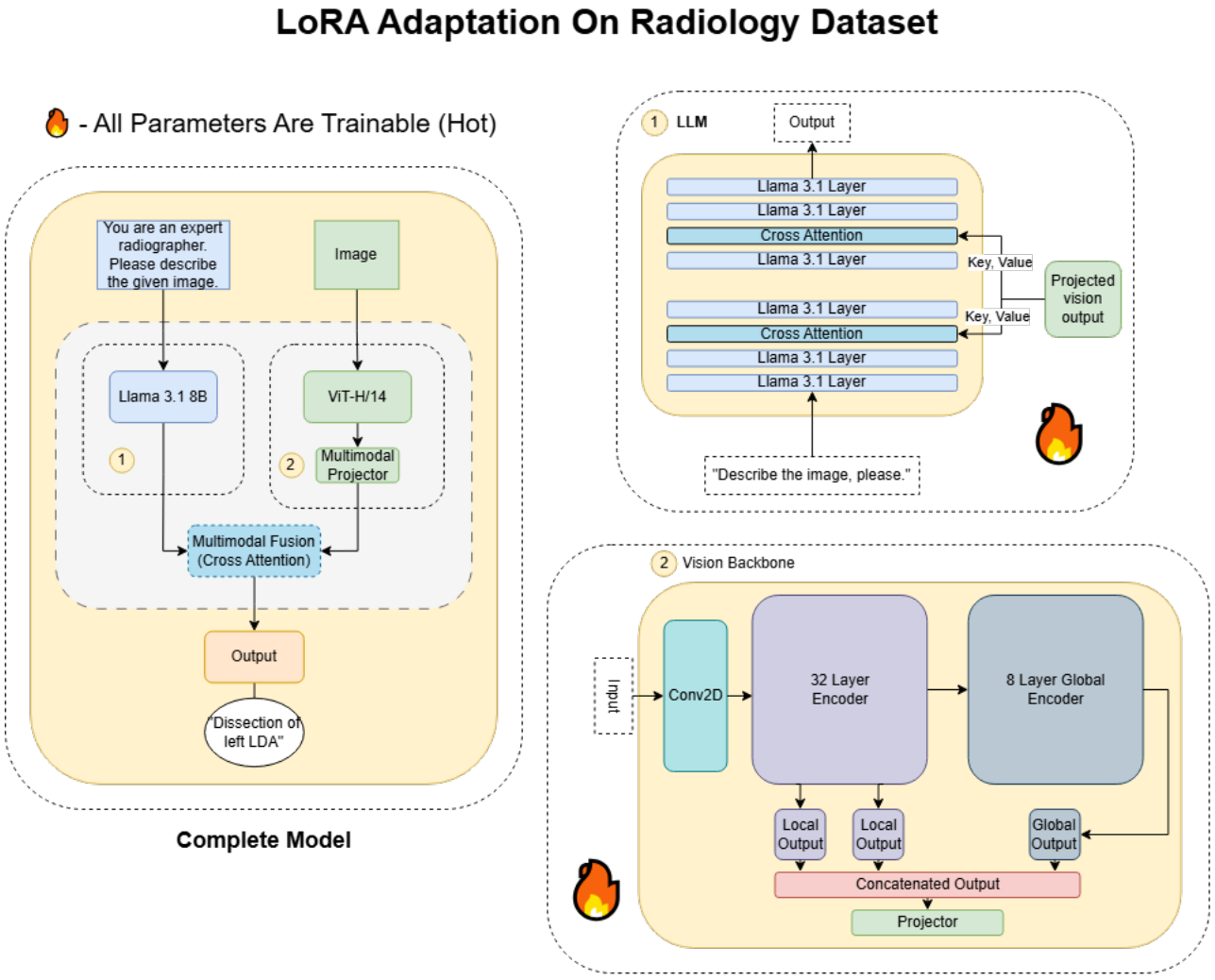


Figure 4.2: LLaMA 3.2-11B Vision-Instruct Architecture

4.2.1 Architecture Overview

The LLaMA 3.2-11B Vision-Instruct model integrates a powerful language model with a vision encoder through a series of cross-attention layers, enabling seamless multimodal understanding.

- **Language Model:** Built upon the LLaMA 3.1 architecture, the language component consists of 40 transformer layers, including 32 self-attention layers and 8 cross-attention layers. This structure allows the model to generate coherent and contextually relevant text based on both textual and visual inputs.

- **Vision Encoder:** The vision module employs a two-stage transformer-based encoder: a 32-layer encoder followed by an 8-layer global encoder. This design captures both local and global visual features, facilitating comprehensive image understanding.
- **Integration Mechanism:** Cross-attention layers serve as the bridge between the vision encoder and the language model. Visual features extracted by the vision encoder are fed into these layers, allowing the language model to attend to relevant visual information during text generation.

4.2.2 Training and Fine-Tuning

The model was pre-trained on a large corpus of image-text pairs, enabling it to learn rich representations across modalities. For domain-specific applications, such as medical imaging, the model can be fine-tuned using parameter-efficient techniques like Low-Rank Adaptation (LoRA). This approach updates a subset of the model's parameters, allowing for efficient adaptation without extensive computational resources.

4.2.3 Applications in Medical Imaging

In the context of medical imaging, the LLaMA 3.2-11B Vision-Instruct model can be fine-tuned on datasets like ROCOv2, which contains radiological images paired with expert annotations. This fine-tuning enables the model to perform tasks such as:

- Generating detailed image captions that describe anatomical structures and potential abnormalities.
- Answering clinical questions based on visual data.
- Assisting in diagnostic processes by providing contextually relevant information derived from medical images.

4.3 Modules Used

Hardware: NVIDIA H100 80GB SXM4 The NVIDIA H100 GPU's ultra-high memory bandwidth and computational power accelerate the training and inference of large-scale models. Its robust architecture ensures seamless execution of memory-intensive tasks.

Framework: Unsloth The Unsloth framework simplifies the implementation of vision-language workflows. It integrates seamlessly with models like Llama Vision 3.2 and supports LoRA adapters for parameter-efficient training. Key features include gradient checkpointing, custom data collators, and modular fine-tuning capabilities.

Optimizer: AdamW with Precision AdamW combines adaptive learning rates with weight decay regularization, ensuring stable and efficient optimization. The use of mixed-precision further reduces memory consumption, enabling large-scale model training on GPUs with limited resources.

Training Framework: PyTorch PyTorch's dynamic computation graph and GPU acceleration make it the ideal framework for developing and fine-tuning complex deep learning models. Its flexibility supports diverse use cases in vision-language processing.

Additional Libraries and Tools:

- **Hugging Face Datasets:** Used for loading and preprocessing the ROCO Radiology dataset.
- **Transformers Library:** Enables seamless model loading, tokenizer usage, and text generation.
- **TextStreamer:** Provides real-time streaming of generated outputs for efficient evaluation.
- **Pillow (PIL):** Facilitates image handling and conversion during dataset processing.
- **NumPy:** Utilized for numerical operations during dataset transformations and model preprocessing.
- **Weights & Biases (WandB):** For logging training metrics and visualizing model performance.

4.4 Parameter Efficient Finetuning with LoRA

We aim to fine-tune the Llama Vision 3.2 11B parameter model on the ROCOv2 radiology dataset to adapt its rich vision-language capabilities for improved interpretation of medical images. To achieve this in a computationally efficient manner, we employ the Low-Rank Adaptation (LoRA) technique. LoRA allows us to update only a small subset of additional parameters rather than the full model weights.

Specifically, instead of fine-tuning the entire weight matrix W_0 from the pre-trained model, we introduce a low-rank update ΔW such that:

$$W = W_0 + \Delta W, \quad \text{with} \quad \Delta W = A \times B$$

Here, $A \in R^{d \times r}$ and $B \in R^{r \times k}$, where the rank r is chosen so that $r \ll \min(d, k)$. This low-rank factorization significantly reduces the number of parameters that need to be fine-tuned, making the process both memory- and compute-efficient.

By updating only the matrices A and B, LoRA enables the model to adapt to the nuances of radiology images while preserving the robust features already learned during pre-training. This targeted approach is particularly beneficial when working with specialized datasets like ROCOv2, where maintaining a balance between domain-specific adaptation and generalization is crucial.

4.5 Mathematical Formulation of Llama Vision

In this section, we provide a streamlined overview of the core components of Llama Vision 3.2, highlighting the key transformations and attention mechanisms that enable powerful multimodal reasoning.

1. Image Patch Embedding.

Let

$$I \in R^{C\times H\times W}$$

A convolutional patch embedder

$$E = \text{Conv2D}(I; W_{\text{conv}}), \quad W_{\text{conv}} \in R^{d\times C\times p\times p},$$

produces feature maps

$$E \in R^{d\times H'\times W'}, \quad H' = \tfrac{H}{p},\ W' = \tfrac{W}{p},$$

which are then flattened into the sequence

$$\widetilde{E} \in R^{N\times d}, \quad N = H'\,W'.$$

2. Class Token & Positional Encoding.

We prepend a learnable class token and add positional embeddings, yielding

$$X = [\,c;\ \widetilde{E}\,] + P \ \in\ R^{(N+1)\times d}\,.$$

3. Vision Encoder (Self-Attention).

Over L layers, residual self-attention refines X:

$$X \ \leftarrow\ X + \text{softmax}\Big(\tfrac{XW_Q\,(XW_K)^\top}{\sqrt{d}} + M\Big)\,XW_V.$$

4. Feature Fusion.

Global context G and selected local features $\{L_i\}$ are concatenated:

$$F = \text{Concat}\big(G, \{L_i\}\big)\ \in\ R^{(N+1)\times d_{\text{fusion}}}, \quad d_{\text{fusion}} = d\,\big(1 + |I|\big).$$

This fused representation is projected into the text-embedding space:

$$F_{\text{proj}} = F\,W_{\text{proj}}, \quad W_{\text{proj}} \in R^{d_{\text{fusion}}\times d_{\text{text}}}.$$

5. Cross-Modal Attention & Output.

Visual context injection into text embeddings T:

$$\text{CrossAttn}(T, F_{\text{proj}}) = \text{softmax}\Big(\tfrac{T\,W_Q^t\,(F_{\text{proj}}\,W_K^i)^\top}{\sqrt{d_{\text{text}}}} + M'\Big)\,(F_{\text{proj}}\,W_V^i).$$

Final normalization and linear head produce logits:

$$\hat{T} = \text{LayerNorm}\big(T + \text{CrossAttn}(T, F_{\text{proj}})\big), \quad \text{Logits} = \hat{T}\,W_{\text{LM}}.$$

Chapter 5

REQUIREMENTS

5.1 Hardware Requirements:

- **Processing Units**
  - **GPU:**
    - NVIDIA A100 or H100 (minimum 80 GB VRAM): For efficient training of large-scale models like LLaMA Vision.
    - CUDA-enabled GPUs are essential for acceleration.
  - **CPU:**
    - Intel Core i9 or AMD Ryzen 9 (minimum 8 cores and 16 threads): For preprocessing and lightweight inference tasks.
- **Memory (RAM)**
  - 64 GB or higher: For handling large datasets and parallel data loading during training.
- **Storage**
  - **SSD:**
    - Minimum 2 TB NVMe SSD: For fast access to datasets and intermediate files during training.
  - **HDD:**
    - Minimum 4 TB: For archiving datasets and pre-trained models.
- **Networking**
  - High-speed internet (1 Gbps or higher): For downloading datasets, model weights (e.g., LLaMA Vision checkpoints), and interacting with cloud services.

5.2 Software Requirements:

- **Programming Languages**
  - Python 3.8 or later: For model development, data preprocessing, and scripting.
- **Frameworks and Libraries**
  - PyTorch (latest stable version): For developing and fine-tuning the vision-language model.
  - Transformers library (Hugging Face): For using and adapting pre-trained models like LLaMA Vision, CLIP, or BLIP.
  - NumPy and pandas: For data handling and preprocessing.
  - OpenCV and PIL: For image manipulation and augmentation.
  - scikit-learn: For evaluation metrics and basic preprocessing.
  - Weights & Biases (wandb): For tracking experiments, model metrics, and visualization.
- **Environment and Tools**
  - Jupyter Notebook: For interactive development and experimentation.
  - Docker: For containerization and portability of the application.
  - Git and GitHub: For version control and collaborative development.
- **APIs and Cloud Services**
  - AWS S3 or Azure Blob Storage: For storing datasets and model checkpoints.
  - Azure DevOps: For CI/CD pipeline setup.
- **Data Requirements**
  - Pre-trained Vision-Language Models: Access to LLaMA Vision or similar models for initial experiments.
  - Datasets: Image-text datasets such as MS COCO, Open Images, or domain-specific datasets like ROCOv2-Radiology, PathVQA, etc
- **Development Environment**
  - **Operating System:** Linux (Ubuntu 20.04 or later) or Windows 11 with WSL2 (Windows Subsystem for Linux).
  - **Virtual Environment:** Anaconda or Miniconda: For managing dependencies and isolating environments.

- **Cloud Services (Optional)**
  - **Compute Instances:**
    - AWS EC2 with GPU support (e.g., p3.16xlarge) or Azure Virtual Machines with NC series to run LLaMA Vision efficiently.
  - **Storage:** Cloud storage solutions like Google Cloud Storage, AWS S3, or Azure Blob Storage.

Chapter 6

EXPERIMENTATION

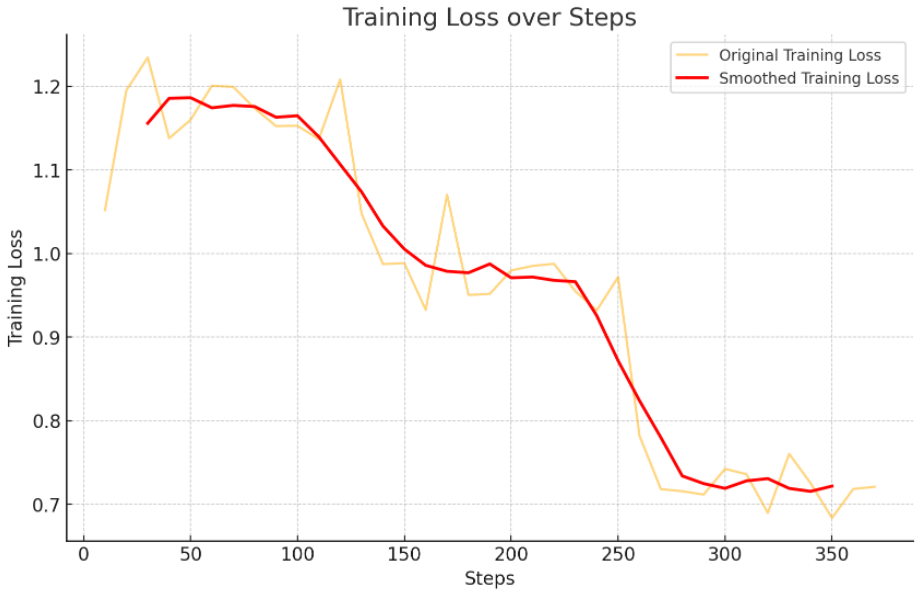


Figure 6.1: Training Loss Over Steps

In our experimentation phase, we initially benchmarked performance using the **PaliGemma model**, a lightweight baseline vision-language model trained on general image-text pairs. However, the PaliGemma model lacked the granularity and clinical precision necessary for interpreting radiology reports and imaging data, resulting in suboptimal diagnostic outputs when applied to specialized modalities such as panoramic radiographs and CT scans.

To address this, we fine-tuned a more advanced vision-language architecture using a curated radiology dataset composed of annotated X-Ray, CT, and ultrasound images. Each sample was paired with professionally written captions detailing observed medical conditions, offering valuable supervised signal for domain adaptation.

For fine-tuning, we adopted **Low-Rank Adaptation (LoRA)**, a parameter-efficient training method that allows effective fine-tuning of large-scale pre-trained models without updating all parameters. In our setup, only **1.22%** of the total model parameters were fine-tuned. LoRA was configured as shown in Table 6.1, enabling selective adaptation of *MLP components, cross-attention layers, and relevant vision-language layers*, significantly reducing the computational burden while preserving fine-tuning efficiency.

Table 6.1: LoRA Configuration Parameters

Parameter	**Value**	**Description**
Dropout rate	0.1	Probability of dropping units to prevent overfitting.
Alpha	32	Scaling factor for low-rank weight updates.
Rank	32	Dimensionality of the low-rank adaptation matrices.

Training was conducted using Hugging Face's **SFTTrainer** with the hyperparameters listed in Table 6.2.

Table 6.2: Training Hyperparameters

Hyperparameter	**Setting**	**Description**
Optimizer	AdamW (8-bit)	Weight-decay optimizer with 8-bit quantization for memory efficiency.
Scheduler	Linear learning rate	Linearly decays the learning rate from initial value to zero.
Learning rate	1.5×10^{-4}	Initial step size for gradient updates.
Batch size (per device)	8	Number of samples processed per GPU before updating gradients.
Gradient accumulation	2 steps	Number of forward passes to accumulate gradients before each update.
Warm-up steps	10	Steps to gradually increase the learning rate at the start of training.
Epochs	3 (372 steps total)	Number of full passes over the training dataset (total update steps).

Training was performed on an **NVIDIA H100 GPU (80 GB)**, with a peak memory usage of **54.6 GB** (69.1 % of total capacity).

Prior to fine-tuning, the model struggled with identifying complex radiographic patterns. After LoRA-based adaptation, performance significantly improved. For instance, the model was able to accurately describe features like an *osteolytic lesion with sinus floor resorption* in panoramic imaging, closely matching expert annotations and showcasing enhanced visual–semantic reasoning.

Training Analysis:

Figure 6.1 visualizes the training loss over steps. The yellow line represents the *original loss*, which shows substantial noise in the early training stages as the model adjusts to domain-specific inputs. The red curve shows the *smoothed loss*, which clearly highlights a downward trend beginning around step 150. This phase marks the point where the model starts to meaningfully learn the mappings between medical imagery and textual annotations. Toward the end of training, the loss stabilizes around **0.7**, indicating convergence and successful optimization of key parameters.

Chapter 7

RESULT AND ANALYSIS

7.1 Evaluation of Fine-Tuned Llama Vision Model

In our evaluation of the fine-tuned Llama Vision model, the generated captions were relatively accurate and closely aligned with expert-provided descriptions. For instance, when presented with a PA chest radiograph (Figure 7.1), the actual caption stated: *"PA chest radiograph reveals a right hilar mass."* The model generated a detailed and specific description: *"Plain chest X-ray in postero-anterior (PA) projection. A well-defined, large, homogeneous mass in the left superior hilum is seen."* While this output identified the mass accurately, it misidentified the side (left vs. right), indicating a challenge in discerning correct laterality.

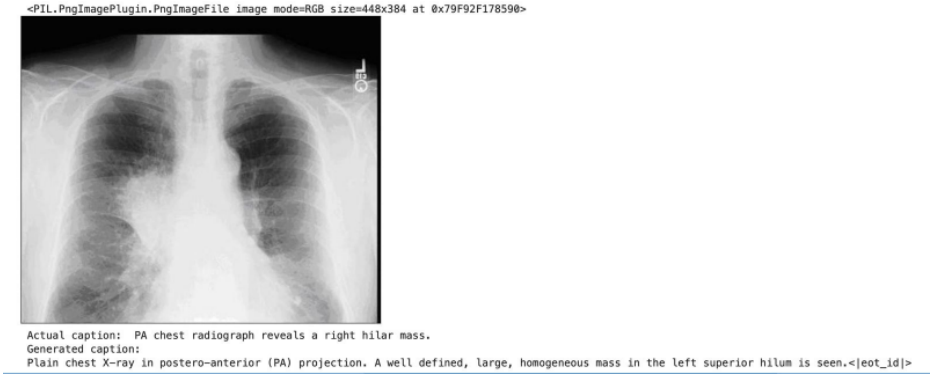
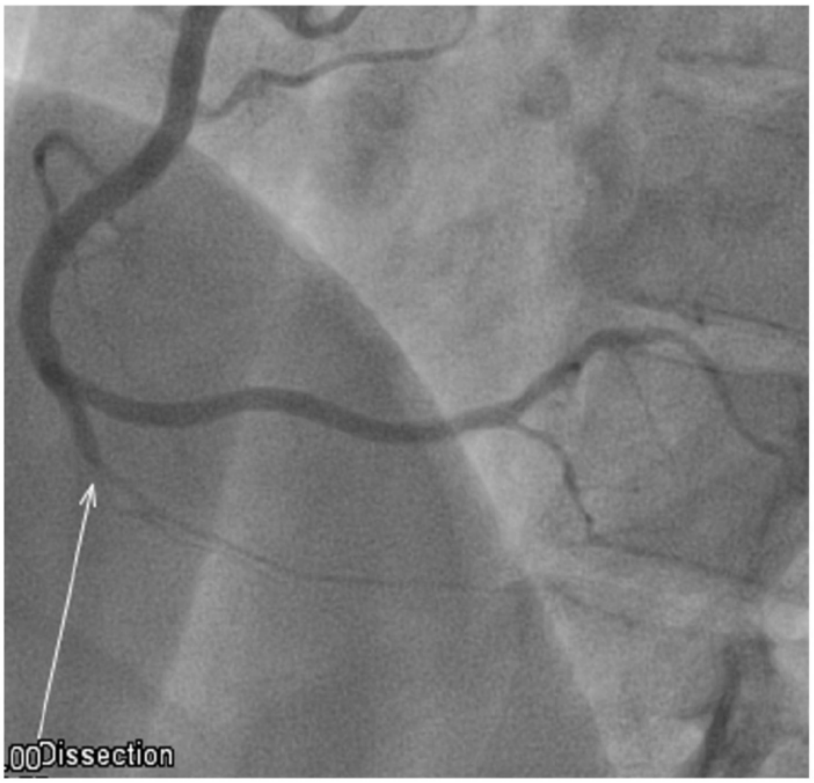


Figure 7.1: PA Chest Radiograph

Similarly, when given an image of a coronary angiogram showing a right descending coronary artery dissection (Figure 7.2), the model's response described a dissection in the left descending coronary artery. Although the side was incorrect, the generated caption still conveyed the correct clinical finding of arterial dissection, demonstrating the model's ability to identify pathology despite laterality confusion.

<PIL.PngImagePlugin.PngImageFile image mode=RGB size=752x725 at 0x79F954F75A10>



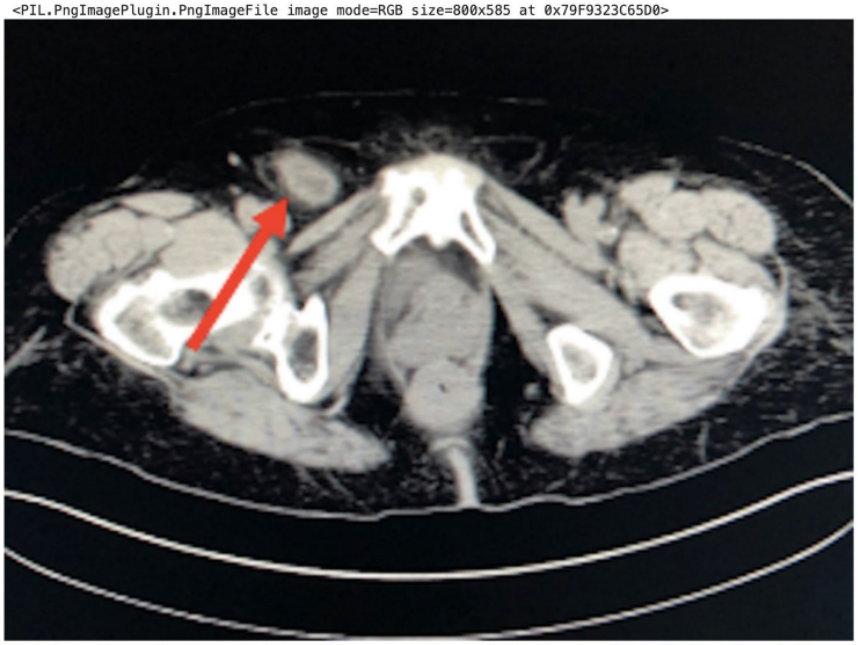
Actual caption: Coronary angiogram showing right descending coronary artery dissection.
Generated caption:
Dissection in left descending coronary artery.<|eot_id|>

Figure 7.2: Coronary Angiogram

As demonstrated by these examples, the model is impressively accurate in generating correct descriptions of the underlying pathology, yet it can struggle to consistently discern correct laterality (*e.g.*, left vs. right).

When analyzing a CT abdomen/pelvis axial view (Figure 7.3), where the actual caption described the presence of a herniated bladder diverticulum, the model generated: *"Computed tomography image of the lower pelvis. Red arrow indicating the herniated bladder."* This output accurately identified the herniated bladder and referenced the arrow in the image, showcasing the model's strong capability in medical imaging interpretation. Overall, the fine-tuned Llama Vision model demonstrated significant progress in producing highly relevant and accurate medical descriptions.

<PIL.PngImagePlugin.PngImageFile image mode=RGB size=800x585 at 0x79F9323C65D0>



Actual caption: CT abdomen/pelvis axial view: the arrow indicating herniated bladder diverticulum
Generated caption:
Computed tomography image of the lower pelvis. Red arrow indicating the herniated bladder.<|eot_id|>

Figure 7.3: CT Abdomen/Pelvis Axial View

However, challenges remain in ensuring consistent outputs, particularly regarding correct laterality. Additionally, the model may misinterpret or overlook critical diagnostic details. For instance, in Figure 7.4, the model "hallucinated" content about "brain-tissue outlines with white lines" and "interhemispheric pressure," even though the original caption describes dysgenesis of the corpus callosum. These shortcomings highlight the need for continued refinement through additional domain-specific training, leveraging larger and more diverse multimodal datasets, and potentially integrating ensemble techniques to improve robustness and diagnostic accuracy.

<PIL.PngImagePlugin.PngImageFile image mode=RGB size=414x453 at 0x79F937D1F350>

Actual caption: Sagittal slice of T1-weighted image showing dysgenesis of the corpus callosum (arrowheads).
Generated caption:
MRI scan (sagittal slice) of patient #1. Brain tissue is outlined with white lines. The orange region is the interhemispheric fissure.<|eot_id|>

Figure 7.4: Sagittal Slice of T1-Weighted Image

7.2 Metrics based Evaluation

The evaluation of the model performance is conducted using METEOR, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1 metrics, with the following observations:

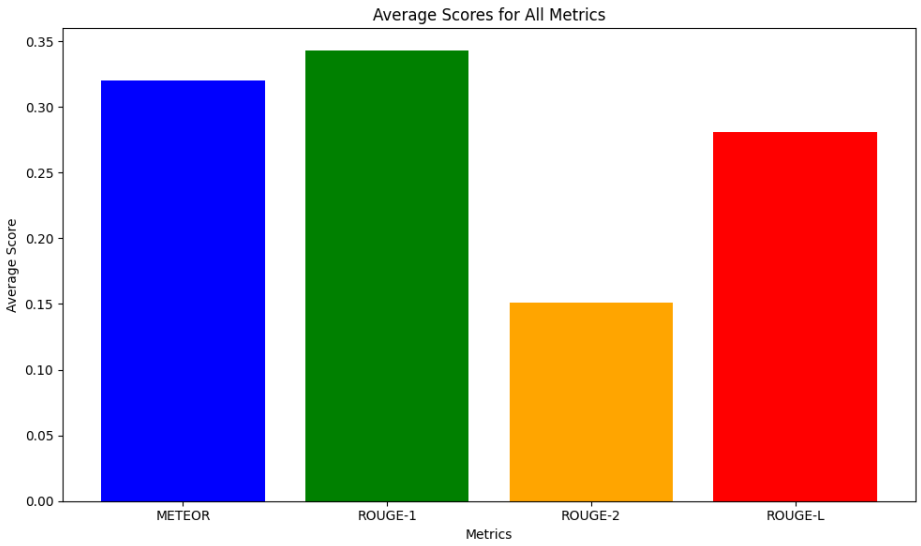


Figure 7.5: Evaluation using ROUGE and METEOR

ROUGE and METEOR Scores: The scores for METEOR (0.32), ROUGE-1 (0.34), and ROUGE-L (0.27) (Figure 7.5) indicate a decent overlap between the generated outputs and the reference texts. However, the comparatively low ROUGE-2 score (0.15) highlights that the model struggles with capturing precise bigram-level overlap, which is indicative of challenges in exact phrase reproduction.

These metrics (ROUGE and METEOR) primarily focus on surface-level word overlap and structural matching, which do not fully account for semantic similarity. This is a limitation when evaluating models that generate text with similar meaning but slightly different word choices. As a result, these scores may undervalue the true quality of the outputs, especially for tasks requiring a deeper semantic understanding.

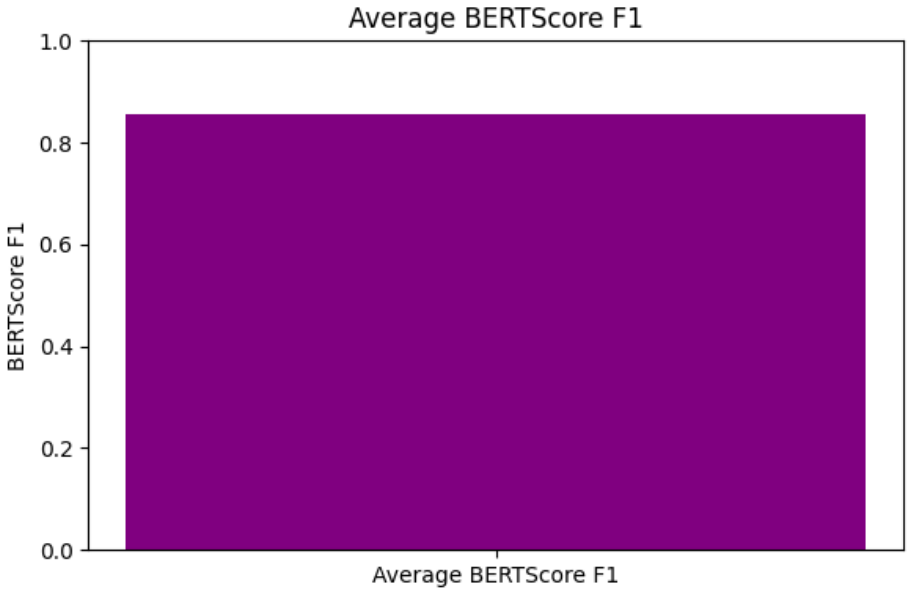


Figure 7.6: Evaluation using BERTSCORE F1

BERTScore F1: The BERTScore F1 (0.85) (Figure 7.6) is notably higher, reflecting a stronger semantic similarity between the generated and reference outputs. Unlike ROUGE and METEOR, BERTScore leverages contextual embeddings to compare meaning rather than exact words or n-grams. The high score indicates that while the outputs may differ in exact wording, they successfully preserve the intended meaning. **G-Eval** is a framework that uses LLM-as-a-judge with chain-of-thoughts (CoT) methodology to evaluate LLM outputs based on custom criteria. We configured this metric to specifically assess the accuracy of image captions. Our model achieved a G-Eval score of 0.75, significantly outperforming the base model's score of 0.27.

The **Hallucination Detection** metric employs LLM-as-a-judge to determine whether outputs contain factually correct information by comparing the generated content against the provided context. This metric is particularly crucial in medical evaluations to ensure outputs are not fabricated by the model. Our model achieved a hallucination score of 0.71, compared to the base model's score of 1.00 (where lower scores indicate fewer hallucinations). Due to the medical nature of our evaluation, we established high threshold values for both metrics to ensure rigorous assessment of model performance.

Table 7.1: Performance Evaluation Metrics

Metric	**Score (Finetuned)**	**Score (Base)**	**Threshold**
GEval	0.75	0.27	0.70
Hallucination	0.71	1.0	0.70
Faithfulness	0.79	0.69	0.70

Chapter 8

CONCLUSION AND FUTURE WORK

In conclusion, our Vision-Language Model (VLM), fine- tuned with LoRA for medical image analysis, has shown notable gains in diagnostic accuracy, precise anatomical refer- ences, and streamlined resource usage. Quantitatively, our approach achieved an average GEval score of 0.75 and a faithfulness score of 0.79, demonstrating significant performance improvements over the base model. Building on these results, we introduced a multi-agent framework organized in a sequential chat model, where each specialized agent—ranging from data extraction to recommendation processes thoroughly analyzed inputs before handing them off. This integrated system yielded exceptional results with a GEval score of 0.898 and perfect faithfulness at 1.000, exceeding our high clinical thresholds.

By integrating the Vision Language Model's refined outputs with orchestrated multi-agent reasoning, the system produces cohesive, context-rich diagnostic reports tailored to each step in the clinical workflow. This integrated approach effectively enhances automated decision support in healthcare, showing substantial promise for improving clinical outcomes through both qualitative improvements and quantitatively validated performance metrics.

8.1 Implications and Future Directions:

8.1.1 Robustness, Bias, and Generalizability

- **Data Diversity:** Conducting model training and validation on geographically and demographically diverse datasets will enhance generalizability and reduce potential biases.

- **Adversarial Testing:** Investigating the model's resilience to adversarial inputs or low-quality images is crucial. Future work will involve stress-testing the model with challenging cases to ensure reliability under real-world conditions.

8.1.2 Educational and Simulation Platforms

- **Synthetic Data Generation:** Leveraging generative capabilities to simulate rare pathologies or complex imaging scenarios can aid in training radiologists and medical students.
- **Interactive Learning Modules:** Developing immersive, AI-driven educational tools—such as question-and-answer interfaces or dynamic case studies—can further enhance medical education and workforce preparedness.

8.2 Future Work

8.2.1 Expanded Clinical Integration

- **Workflow Embedding:** Future research will focus on seamlessly embedding the fine-tuned model within existing clinical workflows, incorporating its outputs into radiology information systems (RIS) and electronic health records (EHRs).
- **User Feedback Loops:** Implementing user feedback mechanisms (e.g., capturing radiologist annotations and corrections in real time) will help continuously refine model accuracy and interpretability.
- **Collaborative Models:** Integrating specialized vision language models with other domain-specific AI models could create an ensemble system, maximizing diagnostic accuracy by cross-checking results from multiple sources (e.g., pathology, genomics).
- **Additional Imaging Modalities:** Extending the model's capabilities to handle modalities like MRI, PET, and mammography would broaden its applicability in complex diagnostic scenarios.

8.2.2 Ensemble Approaches and Multi-Domain Adaptation

- **Collaborative Models:** Integrating specialized vision language models with other domain-specific AI models could create an ensemble system, maximizing diagnostic accuracy by cross-checking results from multiple sources (e.g., pathology, genomics).
- **Additional Imaging Modalities:** Extending the model's capabilities to handle modalities like MRI, PET, and mammography would broaden its applicability in complex diagnostic scenarios.

8.2.3 Regulatory and Ethical Considerations

- **Approval Pathways:** As the vision language model matures, exploring regulatory guidelines (e.g., FDA submissions) for AI-driven medical diagnostics will be essential for real-world deployment.

- **Patient Privacy:** Ongoing research will address data privacy and secure model deployment, ensuring compliance with healthcare regulations (e.g., HIPAA, GDPR).

By addressing these avenues in future work, the project envisions a comprehensive ecosystem where domain-specialized vision language models become integral to routine medical diagnostics and education. These advancements will further bridge the gap between rapidly evolving AI capabilities and the clinical demands of accuracy, interpretability, and efficiency—ultimately elevating patient outcomes and healthcare delivery standards.

Bibliography

[1] Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in healthcare: Applications, challenges, and future outlook (preprint). *Journal of Medical Internet Research*, 26:e59505, August 2024.

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[3] Pieter Thomas Boonen, Nico Buls, Gert Van Gompel, Yannick De Brucker, Dimitri Aerden, Johan De Mey, and Jef Vandemeulebroucke. *Automated Quantification of Blood Flow Velocity from Time-Resolved CT Angiography*, page 11–18. Springer International Publishing, January 2018.

[4] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[6] Dawei Dai, Yuanhui Zhang, Qianlan Yang, Long Xu, Xiaojing Shen, Shuyin Xia, and Guoyin Wang. Pathologyvlm: a large vision-language model for pathology image understanding. *Artificial Intelligence Review*, 58(6), March 2025.

[7] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

[8] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, 7, November 2024.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[10] Ravi Kishore Kodali, Yatendra Prasad Upreti, and Lakshmi Boppana. A quantization approach for the reduced size of large language models. In *2024 16th International Conference on Knowledge and Smart Technology (KST)*, pages 144–148, 2024.

[11] Jiedong Lang, Zhehao Guo, and Shuyu Huang. A comprehensive study on quantization techniques for large language models. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 224–231, 2024.

[12] Jewon Lee, Ki-Ung Song, Seungmin Yang, Donguk Lim, Jaeyeon Kim, Wooksu Shin, Bo-Kyeong Kim, Yong Jae Lee, and Tae-Ho Kim. Efficient llama-3.2-vision by trimming cross-attended visual features. *arXiv*, 2025.

[13] Chunyu Liu, Yixiao Jin, Zhouyu Guan, Tingyao Li, Yiming Qin, Bo Qian, Zehua Jiang, Yilan Wu, Xiangning Wang, Ying Feng Zheng, and Dian Zeng. Visual–language foundation models in medicine. *The Visual Computer*, July 2024.

[14] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12, 2024.

[15] Aawez Mansuri and Judy W. Gichoya. Context is everything: understanding variable llm performance for radiology retrieval-augmented generation. *Radiology Artificial Intelligence*, 7(3), May 2025.

[16] Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024.

[17] Rahul Thapa, Kezhen Chen, Ian Covert, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Dragonfly: Multi-resolution zoom-in encoding enhances vision-language models. *arXiv preprint arXiv:2406.00977*, 2024.

[18] Dane A. Weinert and Andreas M. Rauschecker. Enhancing large language models with retrieval-augmented generation: A radiology-specific approach. *Radiology Artificial Intelligence*, March 2025.

[19] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256, 2023.

[20] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[21] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735, 2025.

[22] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

Github Link

https://github.com/eng21am0117/Team-1-Major-Project

vlmreport

by Vinutha N

Submission date: 22-May-2025 07:38PM (UTC+0530)
Submission ID: 2682224698
File name: VLM___New_Report_Format-2.pdf (1.53M)
Word count: 8390
Character count: 51275

DAYANANDA SAGAR UNIVERSITY

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt,
Bengaluru-562112, Karnataka, India

SCHOOL OF
ENGINEERING

Bachelor of Technology
in
COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

A Project Report On

Enhancing Medical Diagnostics with Vision-Language Models

By
Shriyans Shriniwas Arkal - ENG21AM0117
Sri Bharath Sharma P - ENG22AM3005
Yudhajit Jana - ENG22AM3021

Under the supervision of
Dr. Vinutha N
Associate Professor
Computer Science & Engineering (AI & ML)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

SCHOOL OF ENGINEERING
DAYANANDA SAGAR UNIVERSITY

(2024 – 2025)

DAYANANDA SAGAR UNIVERSITY

SCHOOL OF ENGINEERING

Department of Computer Science & Engineering

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara Dt, Bengaluru-562112, Karnataka, India

CERTIFICATE

This is to certify that the project entitled **"Enhancing Medical Diagnostics with Vision-Language Models"** is carried out by **Shriyans Shriniwas Arkal (ENG21AM0117), Sri Bharath Sharma P (ENG22AM3005), Yudhajit Jana (ENG22AM3021)**, bonafide students of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore, in partial fulfillment for the award of a degree in Bachelor of Technology in Computer Science and Engineering, during the year **2024 - 2025**.

Dr. Vinutha N	**Dr. Vinutha N**	**Dr. Jayavrinda Vrindavanam**
Associate Professor	Project Co-ordinator	Professor & Chairperson
Dept. of CSE (AI&ML)	Dept. of CSE (AI&ML)	Dept. of CSE (AI&ML)
School of Engineering	School of Engineering	School of Engineering
Dayananda Sagar University	Dayananda Sagar University	Dayananda Sagar University
Signature	Signature	Signature .....

Name of the Examiners:	Signature with date:
1.....	.....
2.....	.....
3.....	.....

vlmreport

ORIGINALITY REPORT

19%	12%	10%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Birla Institute of Technology and Science Pilani Student Paper	8%
2	arxiv.org Internet Source	2%
3	Guo, Danfeng. "Applying Medical Language Models to Medical Image Analysis", University of California, Los Angeles, 2024 Publication	1%
4	Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir et al. "ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset", Scientific Data, 2024 Publication	<1%
5	Tianhan Xu, Bin Li. "KELLM: Knowledge-Enhanced Label-Wise Large Language Model for Safe and Interpretable Drug Recommendation", Electronics, 2025 Publication	<1%
6	www.dsu.edu.in Internet Source	<1%
7	ceur-ws.org Internet Source	<1%
8	Submitted to Liverpool John Moores University Student Paper	<1%

9	Wenpin Hou, Qi Liu, Huifang Ma, Yilong Qu, Zhicheng Ji. "Assessing large multimodal models for one-shot learning and interpretability in biomedical image classification", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
10	www.medrxiv.org Internet Source	<1 %
11	www.transparencymarketresearch.com Internet Source	<1 %
12	Qi Hong, Shijie Liu, Liying Wu, Qiqi Lu et al. "Evaluating the performance of large language & visual-language models in cervical cytology screening", Springer Science and Business Media LLC, 2025 Publication	<1 %
13	stars.library.ucf.edu Internet Source	<1 %
14	Pandya, Krutik. "Automated Software Compliance Using Smart Contracts and Large Language Models in Continuous Integration and Continuous Deployment With DevSecOps", Arizona State University, 2024 Publication	<1 %
15	www.medicai.io Internet Source	<1 %
16	Submitted to University of Newcastle upon Tyne Student Paper	<1 %
17	"Proceedings of IEMTRONICS 2024", Springer Science and Business Media LLC, 2025 Publication	<1 %

18	Submitted to University of Sunderland Student Paper	<1%
19	Alexander Novikov. "SKYNET 2023 Conception of the Artificial Super Intelligence Project. A System Approach. ver. 3", Open Science Framework, 2023 Publication	<1%
20	Submitted to British University in Egypt Student Paper	<1%
21	Submitted to Cranfield University Student Paper	<1%
22	Submitted to University of Dundee Student Paper	<1%
23	Submitted to University of Essex Student Paper	<1%
24	realrads.com Internet Source	<1%
25	www.scitepress.org Internet Source	<1%
26	www.tomorrow.bio Internet Source	<1%
27	"Computer Vision – ECCV 2024", Springer Science and Business Media LLC, 2025 Publication	<1%
28	8grams.medium.com Internet Source	<1%
29	Submitted to City University of Hong Kong Student Paper	<1%
30	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1%

31	Submitted to KUMARAGURU COLLEGE OF TECHNOLOGY Student Paper	<1%
32	Madhusudan Ghosh, Shrimon Mukherjee, Asmit Ganguly, Partha Basuchowdhuri, Sudip Kumar Naskar, Debasis Ganguly. "AlpaPICO: Extraction of PICO Frames from Clinical Trial Documents Using LLMs", Methods, 2024 Publication	<1%
33	etd.uum.edu.my Internet Source	<1%
34	ir.lib.uwo.ca Internet Source	<1%
35	"Pattern Recognition", Springer Science and Business Media LLC, 2025 Publication	<1%
36	hdl.handle.net Internet Source	<1%
37	studentshare.org Internet Source	<1%
38	web.media.mit.edu Internet Source	<1%
39	Lawrence K.Q. Yan, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, Junyu Liu. "Large Language Model Benchmarks in Medical Tasks", Open Science Framework, 2024 Publication	<1%
40	howiehwong.github.io Internet Source	<1%
41	job-search.astrazeneca.ru Internet Source	<1%

42	postdicom.com Internet Source	<1 %
43	pubmed.ncbi.nlm.nih.gov Internet Source	<1 %
44	www.openu.ac.il Internet Source	<1 %
45	www.research-collection.ethz.ch Internet Source	<1 %
46	Bisheng Yang, Zhen Dong, Fuxun Liang, Xiaoxin Mi. "Ubiquitous Point Cloud - Theory, Model, and Applications", CRC Press, 2024 Publication	<1 %
47	Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023) - Proceedings of the International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023), Dec 1–2, 2023, Kolkata, India", CRC Press, 2024 Publication	<1 %
48	Zijie Cai, Hui Fang, Jianhua Liu, Ge Xu, Yunfei Long, Yin Guan, Tianci Ke. "Improving unified information extraction in Chinese mental health domain with instruction-tuned LLMs and type-verification component", Artificial Intelligence in Medicine, 2025 Publication	<1 %
49	aclanthology.org Internet Source	<1 %
50	"Computer Vision – ACCV 2024", Springer Science and Business Media LLC, 2025 Publication	<1 %

51	"Intelligent Systems", Springer Science and Business Media LLC, 2025 Publication	<1%
52	Zhao, Fengxiang. "Building an Extensible, AI-Augmented Ecological Momentary Assessment Platform.", University of Missouri - Columbia, 2024 Publication	<1%
53	"Advanced Intelligent Computing Technology and Applications", Springer Science and Business Media LLC, 2024 Publication	<1%
54	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2025 Publication	<1%
55	Mingzhe Hu, Joshua Qian, Shaoyan Pan, Yuheng Li, Richard L J Qiu, Xiaofeng Yang. "Advancing medical imaging with language models: featuring a spotlight on ChatGPT", Physics in Medicine & Biology, 2024 Publication	<1%
56	Trivedi, Dhwani Kirankumar. "Semi-Supervised Fish Detection, Classification, and Length Estimation Using Faster-RCNN and Masked Autoencoders.", Northeastern University Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

DAYANANDA SAGAR UNIVERSITY

Deverakaggalahalli, Harohalli, Kanakapura Rd, Dist. Ramanagara, Karnataka-562112

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

CERTIFICATE
OF PARTICIPATION

THIS CERTIFICATE IS PRESENTED TO :

Yudhajit Jana

In recognition of active participation in the "Tech Spark 2.0" event, organized by the Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning) held on 26th April 2025.

Dr. Jayavrinda Vrindavanam
Professor & Chairperson
DEPARTMENT OF CSE(AI & ML)
SOE, DSU

Dr. Udaya Kumar Reddy K.R.
Dean
SCHOOL OF ENGINEERING
DSU

DAYANANDA SAGAR UNIVERSITY

Deverakaggalahalli, Harohalli, Kanakapura Rd, Dist. Ramanagara, Karnataka-562112

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

CERTIFICATE
OF PARTICIPATION

THIS CERTIFICATE IS PRESENTED TO :

Sri Bharath Sharma Perala

In recognition of active participation in the "Tech Spark 2.0" event, organized by the Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning) held on 26th April 2025.

Dr. Jayavrinda Vrindavanam
Professor & Chairperson
DEPARTMENT OF CSE(AI & ML)
SOE, DSU

Dr. Udaya Kumar Reddy K.R.
Dean
SCHOOL OF ENGINEERING
DSU

DAYANANDA SAGAR UNIVERSITY

Devarakaggalahalli, Harohalli, Kanakapura Rd, Dist. Ramanagara, Karnataka-562112

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

CERTIFICATE
OF PARTICIPATION

THIS CERTIFICATE IS PRESENTED TO :

Shriyans Shriniwas Arkal

In recognition of active participation in the "Tech Spark 2.0" event, organized by the Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning) held on 26th April 2025.

Dr. Jayavrinda Vrindavanam
Professor & Chairperson
DEPARTMENT OF CSE(AI & ML)
SOE, DSU

Dr. Udaya Kumar Reddy K.R.
Dean
SCHOOL OF ENGINEERING
DSU

Gmail

Yudhajit Jana <yudhajit.j12@gmail.com>

2025 IEEE International Conference on Electronics, Computing and Communication Technologies : Submission (2267) has been created.

1 message

Microsoft CMT <email@msr-cmt.org> 28 March 2025 at 14:22
To: yudhajit.j12@gmail.com

Hello,

The following submission has been created.

Track Name: Advances in Healthcare Technologies

Paper ID: 2267

Paper Title: Enhancing Medical Diagnostics with Vision-Language Models and Agents

Abstract:
Commercial vision models work well for general imaging tasks but struggle with radiographic diagnosis because they miss the domain-specific details needed for accuracy. This is a major issue since radiography is essential for revealing complex anatomical structures. This paper addresses these limitations by proposing a novel multimodal approach that adapts the Llama 3.2-11B-Vision model using Low-Rank Adaptation (LoRA) on the ROCOv2: Radiology Objects in COntext Version 2 dataset. Our integrated approach achieves an average GEval score of 0.75 and a faithfulness score of 0.79 showing significant performance gains over the base model. Furthermore, to overcome the diagnostic shortcomings of standalone vision-language systems, we introduce an end-to-end multi-agent framework that ingests the diagnostic output of the fine-tuned vision-language model, then uses chain-of-thought reasoning and sequential processing to perform specialized tasks such as patient history extraction, case data interpretation, lab report analysis, summarization, and diagnosis. Our evaluation using DeepEval metrics shows a GEval score of 0.898, and perfect faithfulness at 1.000—exceeding our high clinical thresholds. This integrated approach enhances overall diagnostic performance, bridging the gap between general-purpose vision models and the specialized requirements of accurate radiographic diagnosis.

Created on: Fri, 28 Mar 2025 08:52:48 GMT

Last Modified: Fri, 28 Mar 2025 08:52:48 GMT

Authors:
- yudhajit.j12@gmail.com (Primary)
- bharathperala8@gmail.com
- shriyans.arkal07@gmail.com
- saylibande19@gmail.com
- ratan152004@gmail.com
- vinutha.n-aiml@dsu.edu.in

Secondary Subject Areas: Not Entered

Submission Files:
IEEE_Connect_2025.pdf (538 Kb, Fri, 28 Mar 2025 08:46:14 GMT)

Submission Questions Response: Not Entered

Thanks,
CMT team.

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our Privacy Statement.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Enhancing Medical Diagnostics with Vision-Language Models and Agents

1st Sri Bharath Sharma Perala
Department of CSE (AI & ML)
Dayananda Sagar University
Bangalore, India
bharathperala8@gmail.com

2nd Shriyans Shriniwas Arkal
Department of CSE (AI & ML)
Dayananda Sagar University
Bangalore, India
shriyans.arkal07@gmail.com

3rd Sayli Pankaj Bande
Department of CSE (AI & ML)
Dayananda Sagar University
Bangalore, India
saylibande19@gmail.com

4th Ratan Ravichandran
Department of CSE(AI & ML)
Dayananda Sagar University
Bangalore, India
ratan152004@gmail.com

5th Yudhajit Jana
Department of CSE (AI & ML)
Dayananda Sagar University
Bangalore, India
yudhajit.j12@gmail.com

6th Vinutha N
Department of CSE (AI & ML)
Dayananda Sagar University
Bangalore, India
vinutha.n-aiml@dsu.edu.in

***Abstract*—Commercial vision models work well for general imaging tasks but struggle with radiographic diagnosis because they miss the domain-specific details needed for accuracy. This is a major issue since radiography is essential for revealing complex anatomical structures. This paper addresses these limitations by proposing a novel multimodal approach that adapts the Llama 3.2-11B-Vision model using Low-Rank Adaptation (LoRA) on the ROCOv2: Radiology Objects in COntext Version 2 dataset. Our integrated approach achieves an average GEval score of 0.75 and a faithfulness score of 0.79 showing significant performance gains over the base model. Furthermore, to overcome the diagnostic shortcomings of standalone vision-language systems, we introduce an end-to-end multi-agent framework that ingests the diagnostic output of the fine-tuned vision-language model, then uses chain-of-thought reasoning and sequential processing to perform specialized tasks such as patient history extraction, case data interpretation, lab report analysis, summarization, and diagnosis. Our evaluation using DeepEval metrics shows a GEval score of 0.898, and perfect faithfulness at 1.000—exceeding our high clinical thresholds. This integrated approach enhances overall diagnostic performance, bridging the gap between general-purpose vision models and the specialized requirements of accurate radiographic diagnosis.**

***Index Terms*—Vision models, Radiographic diagnosis, Low-Rank Adaptation (LoRA), Chain-of-thought reasoning**

I. Introduction

Radiology is pivotal for diagnosis and treatment planning, yet increasing data complexity—from high-resolution CT, MRI, and X-ray images to comprehensive clinical cases—has exposed significant shortcomings in traditional diagnostic workflows. Existing systems rely on single-agent or monomodal approaches that inadequately integrate visual and textual information, leading to suboptimal diagnostic accuracy and efficiency. Our project tackles these challenges by fine-tuning a state-of-the-art vision-language model using techniques such as Low-Rank Adaptation (LoRA) and embedding it within a multi-agent AI framework. This framework leverages specialized agents—each designed to handle discrete tasks like patient history extraction, lab data interpretation, and imaging correlation—allowing for chain-of-thought reasoning and cross-validation that enhances both interpretability and robustness. Unlike conventional models that process diverse inputs in a single flow, our multi-agent system distributes the workload among task-specific agents, thereby reducing oversimplification and error propagation commonly seen in current approaches. This targeted division of labor not only improves diagnostic accuracy in complex clinical scenarios but also offers technical advantages such as efficient parameter tuning, dynamic adaptation to new data distributions, and enhanced performance in outlier cases. Ultimately, our approach demonstrates clear improvements over existing methods, paving the way for more reliable and scalable clinical applications.

II. Literature Review

Ayaz et al. [1] introduce MedVLM, a vision-language model integrating Florence-2 and LLaMA-2 via LoRA, optimized for medical applications such as Visual Question Answering (VQA) and medical report generation. Evaluations on Rad-VQA show superior accuracy over specialized and generalist models, with radiologists validating 74% of the generated reports as high quality.

Rückert et al. [2] present ROCOv2 (Radiology Objects in COntext Version 2), a multimodal dataset comprising of 79,789 radiological images with curated captions and medical concepts. It adds 35,705 new images over the previous dataset (ROCO) published in 2018 and supports training image annotation models based on image-caption pairs, or for multi-label image classification.

Hu et al. [3] introduce a novel approach called LoRA (Low-Rank Adaptation of Large Language Models), which adapts pre-trained models by injecting low-rank matrices into fixed weights. This significantly reduces trainable parameters and GPU memory usage, allowing efficient fine-tuning of large-scale models (e.g., GPT-3 175B) with comparable or improved performance.

[4]AutoGen presents a multi-agent framework that leverages question answering models that use RAG to enable context aware interactions, which is better than traditional systems in efficiency and complex decision-making tasks. Alternatively, [5]Agent Hospital introduces a simulacrum-based approach to medical AI training through virtual patient interactions, with MedAgent-Zero allowing doctor agents to evolve across 32 departments and 339 diseases, achieving superior performance on the MedQA benchmark and accelerating training without heavy manual data annotation.

III. Methodology

A. Dataset

1) Vision Language Model Dataset: We utilize the Radiology Objects in COntext version 2 (ROCOv2) dataset for fine-tuning. ROCOv2 is a comprehensive multimodal repository comprising 79,789 radiological images paired with corresponding medical concepts and captions extracted from the PMC Open Access subset. For our fine-tuning experiments, we sample 1,978 representative examples from ROCOv2. To align with the input requirements of our vision-language model, each sample is transformed into a structured conversational format. Specifically, each data instance is reformulated into a dialogue consisting of a user prompt and an assistant response. The user prompt includes both a textual instruction and the corresponding image, while the assistant response contains the associated caption. An example of this structure is provided below:

```
[
  {
    "role": "user",
    "content": [
      {"type": "text", "text": "You are an
      ↪  expert radiographer. Describe
      ↪  accurately what you see in this
      ↪  image."},
      {"type": "image", "image":
      ↪  "sample[""image""]"}
    ]
  },
  {
    "role": "assistant",
    "content": [
      {"type": "text", "text":
      ↪  "sample[""caption""]"}
    ]
  }
]
```

Fig. 1: Dataset Instance for ROCOv2

2) Agent Dataset: The agent framework integrates three key data sources to support a radiological decision system. Patient history from structured EHRs, stored for extraction by the Patient Profile & History agent. Clinical case files, maintained as unstructured text, are processed to extract diagnostic cues via the Case Data Extractor, while lab reports detailing both standard tests and specialized assessments are managed by the Lab Report Analyzer. All sources feed into RAG pipelines that supply each agent with high-quality, context-specific information.

B. Hybrid Framework of Vision Language Model with Agentic Framework

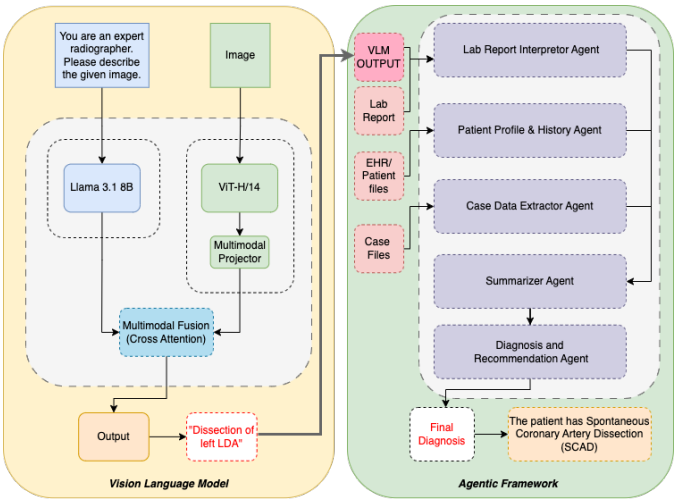


Fig. 2: Diagnostic Framework for Spontaneous Coronary Artery Dissection

Illustrated in 'Fig. 2', our proposed framework consists of two main components: a specialized vision-language model and an advanced multi-agent workflow system. The vision component utilizes Llama 3.2 Vision (11B), fine-tuned on a curated subset of the ROCOv2 dataset, enabling efficient adaptation to domain-specific medical contexts, particularly radiology. This model bridges visual and linguistic semantic spaces, generating clinically relevant diagnostic insights from medical imaging data through natural language processing capabilities. The system's output is then processed through a multi-agent workflow leveraging the Autogen Framework. This workflow comprises tools, External Data inputs relevant to the case, and five key agents: the Patient Profile & History Agent, the Case Data Extractor Agent, the Radiology Lab Report Agent, the Summarizer Agent, and the Diagnosis and Recommendation Agent

1) Vision Language Model Description: We propose a fine-tuned version of the Llama 3.2-11B-Vision model tailored for radiology applications. The Llama model integrates a two-stage vision encoder based on the ViT-H/14 architecture 'Fig. 2', a language model, and a multimodal fusion mechanism. The vision encoder first employs a 32-layer transformer to extract fine-grained local features from image patches, then uses an 8-layer global encoder with gated attention to aggregate contextual information across the image; intermediate features from the local stage are concatenated with the global output to form a rich, multi-scale representation. The language model, based on the Llama 3.1 architecture, is a 40-layer decoder-only transformer that incorporates cross-attention layers every fifth layer, enabling it to integrate visual

features during text generation. The multimodal fusion module projects the concatenated visual features into the language model, effectively aligning the visual and textual semantic spaces. This design captures both detailed local structures and overall image context, facilitating the generation of accurate and contextually nuanced descriptions.

2) Multi-Agent Description: We employ a multi-agent framework by decomposing complex clinical tasks into different components. For our application, AutoGen supports diverse conversational patterns, including sequential, nested, and group interactions. Its design incorporates features such as chat termination, human-in-the-loop integration, and customizable agent behaviors, which are critical for controlled and accurate outputs which are integral in the medical field. We define our multi agent framework as follows; **The Patient Profile Extractor Agent** analyzes patient text data to generate structured historical and demographic views, while the **Case Data Extractor Agent** processes unstructured clinical case files to capture diagnostic observations. Concurrently, the **Lab Reporter Interpreter Agent** evaluates lab results for critical values that correlate with radiographic findings. These outputs are then integrated by the Radiology Summarizer Agent, which compiles a unified overview, and finally, the Diagnosis and Recommendation Agent generates a comprehensive diagnostic report with primary and alternative diagnoses, detailed reasoning, and actionable recommendations.

IV. Experimentation

A. Vision Language Model Experimentation

We harness the capabilities of Unsloth to fine-tune the Llama-3.2-11B-Vision-Instruct model in 16-bit LoRA, concentrating on parameter-efficient updates across both the vision and language components. By training only the LoRA adapters, we significantly reduce computational overhead while still allowing the model to learn task-specific nuances. The fine-tuning procedure targets vision-specific layers, language layers, attention modules, and MLP modules, ensuring a comprehensive adaptation to medical imaging tasks. We utilise the following LoRa specific hyperparameters in Table I

TABLE I: LoRA Hyperparameters

Param	Value	Description
Rank	32	Dimensionality of low-rank adaptation matrix
Alpha	32	Scaling factor for LoRA updates with pretrained weights
Dropout	0.1	Regularization parameter to prevent overfitting

TABLE II: Optimization Strategies for Vision Language Models

Parameter	Value	Description
AdamW 8-bit optimizer	Enabled	Memory-efficient variant of Adam optimizer reducing overhead while preserving updates
Linear learning rate scheduler	Enabled	Applies gradual decay in learning rate to ensure smooth convergence
Gradient checkpointing	Enabled	Recomputes intermediate activations to reduce GPU memory consumption

TABLE III: Training Hyperparameters for Vision Language Models

Parameter	Value	Description
Batch size	8	Number of samples processed in each training step
Gradient accumulation steps	2	Steps to accumulate gradients before weight update
Learning rate	1.5e-4	Step size for parameter updates during optimization
Training epochs	3	Complete passes through the entire dataset

Furthermore, to ensure stable and effective fine-tuning, we have selected a set of hyperparameters and optimization strategies that balance performance with resource constraints mentioned in Table. II and Table. III

B. Agentic Workflows Experimentation

In our experimentation, we evaluated multiple conversational workflows—including non-sequential, round-robin, and group chat models—to determine the most effective approach for our multi-agent diagnostic system. While non-sequential and group chat configurations had lower response times, they often led to information overload, race conditions, and partial context propagation, which can be detrimental in critical medical cases. Round-robin workflows, though evenly distributing agent contributions, proved inefficient when some agents had nothing new to add. Ultimately, the **sequential chat model** proved to be the optimal solution, as its structured, turn-based approach guarantees that each agent's output is fully finalized and contextually complete before the next agent processes the information—a critical requirement to avoid misdiagnosis in the clinical setting.

V. Results

A. Vision Language Model Results

The fine-tuning process was conducted on an NVIDIA H100 80GB GPU for three epochs (372 steps). By restricting training to approximately 1–2% of the parameters using LoRA, runtime per epoch was reduced compared to full-parameter training.

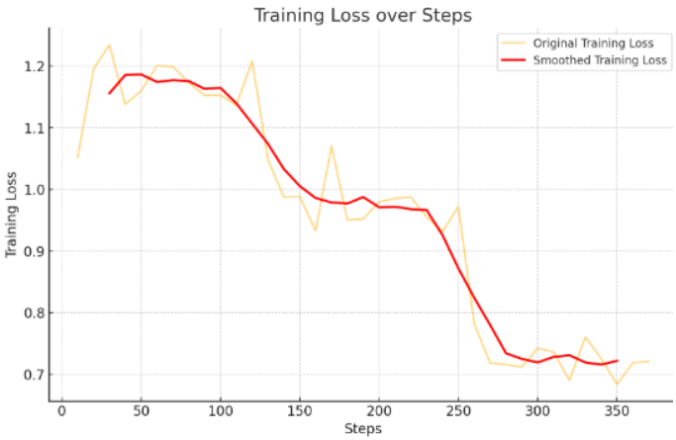


Fig. 3: Training Loss for Vision Model using LoR

Peak GPU utilization was around 54.6 GB (69% of 80GB), with mixed-precision operations (BF16) further lowering memory consumption. Only the newly introduced low-rank weight matrices and select attention/MLP layers were trainable, preserving the core model capacity while enhancing memory efficiency.

Convergence behavior was tracked through this training-loss curve, which initially exhibited some variability but stabilized around 0.7 by the final epoch as show in 'Fig. 3. The steady decline in the smoothed loss metric indicated that updates to the model parameters were improving alignment with expert-generated captions. Notably, early fluctuations were largely mitigated by a brief warmup phase, ensuring more stable optimization in subsequent epochs. As for the model performance, the fine-tuned model displayed significantly improved performance as compared to the baseline model.

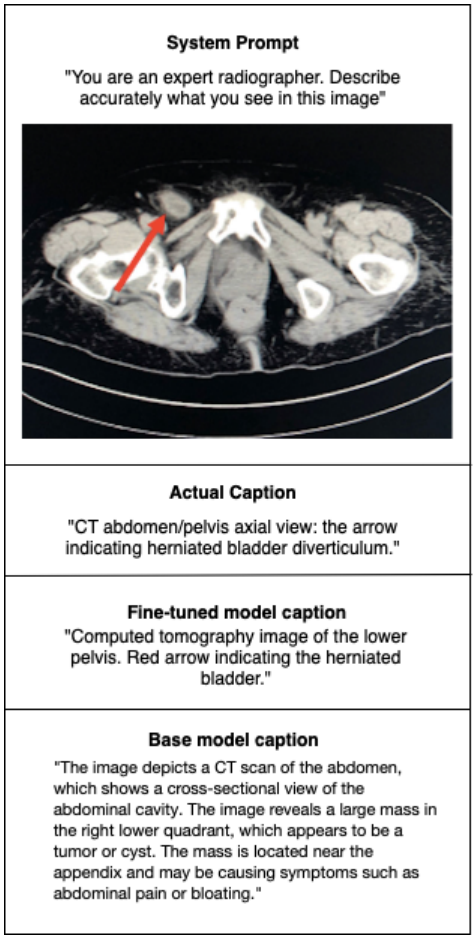


Fig. 4: Vision Language Model Output

Our model demonstrates a clear advantage in medical terminology usage. Specifically, the caption,

"Computed tomography image of the lower pelvis. Red arrow indicating the herniated bladder."

accurately identifies the imaging modality (CT) and locates the pathology (the lower pelvis), while also correctly describing the key finding as a *"herniated bladder."* This aligns closely with the reference caption, which mentions a *"herniated bladder diverticulum"*. Although our model does not capture the full specificity of the reference caption, it provides a significantly more precise and clinically relevant description than the baseline model. By correctly identifying the presence of a herniation and avoiding speculative errors, the model's output demonstrates a strong grasp of medical language and radiological context, bringing it much closer to expert-level interpretation.

When comparing this performance with the baseline model, the difference in clinical accuracy becomes even more apparent. The baseline-generated caption lacks both anatomical precision and correct pathology identification. Unlike our model, which closely aligns with the reference caption, the baseline output introduces an entirely incorrect interpretation by misidentifying the abnormality as a *"tumor or cyst"* rather than a *"herniated bladder diverticulum"*. Additionally, its broad description of the abdomen instead of the lower pelvis reduces its clinical relevance. This mischaracterization highlights the limitations of general vision-language models in medical imaging and underscores the importance of fine-tuning for improved diagnostic accuracy.

B. Evaluation Metrics

For evaluating our Vision Language Model, we employed specialized metrics from the DeepEval framework rather than traditional NLP evaluation techniques such as BERT, ROUGE, or BLEU scores. As shown in Table IV, our evaluation focused on the following metrics. **G-Eval** is a framework that uses LLM-as-a-judge with chain-of-thoughts (CoT) methodology to evaluate LLM outputs based on custom criteria. We configured this metric to specifically assess the accuracy of image captions. Our model achieved a G-Eval score of 0.75, significantly outperforming the base model's score of 0.27. The **Hallucination Detection** metric employs LLM-as-a-judge to determine whether outputs contain factually correct information by comparing the generated content against the provided context. This metric is particularly crucial in medical evaluations to ensure outputs are not fabricated by the model. Our model achieved a hallucination score of 0.71, compared to the base model's score of 1.00 (where lower scores indicate fewer hallucinations). Due to the medical nature of our evaluation, we established high threshold values for both metrics to ensure rigorous assessment of model performance.

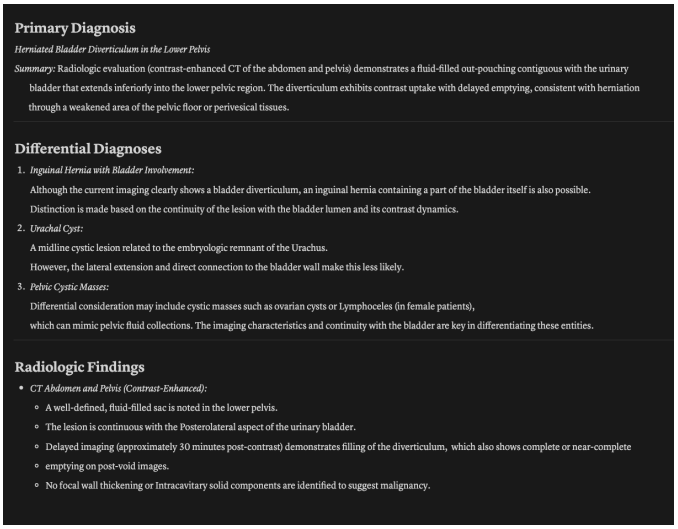
TABLE IV: Performance Evaluation Metrics

Metric	Score (Finetuned)	Score (Base)	Threshold
GEval	0.75	0.27	0.70
Hallucination	0.71	1.0	0.70
Faithfulness	0.79	0.69	0.70

C. Agentic Results

Below is an example of the output from the agentic diagnostic workflow for a case of Herniated Bladder. In this workflow, the visual language model (VLM) generated a

detailed description of the pelvic radiology scan, which was subsequently forwarded to the radiology lab report analyzer agent for further evaluation.



Primary Diagnosis

Herniated Bladder Diverticulum in the Lower Pelvis

Summary: Radiologic evaluation (contrast-enhanced CT of the abdomen and pelvis) demonstrates a fluid-filled out-pouching contiguous with the urinary bladder that extends inferiorly into the lower pelvic region. The diverticulum exhibits contrast uptake with delayed emptying, consistent with herniation through a weakened area of the pelvic floor or perivesical tissues.

Differential Diagnoses

1. *Inguinal Hernia with Bladder Involvement:*
   Although the current imaging clearly shows a bladder diverticulum, an inguinal hernia containing a part of the bladder itself is also possible.
   Distinction is made based on the continuity of the lesion with the bladder lumen and its contrast dynamics.
2. *Urachal Cyst:*
   A midline cystic lesion related to the embryologic remnant of the Urachus.
   However, the lateral extension and direct connection to the bladder wall make this less likely.
3. *Pelvic Cystic Masses:*
   Differential consideration may include cystic masses such as ovarian cysts or Lymphoceles (in female patients),
   which can mimic pelvic fluid collections. The imaging characteristics and continuity with the bladder are key in differentiating these entities.

Radiologic Findings

- *CT Abdomen and Pelvis (Contrast-Enhanced):*
  - A well-defined, fluid-filled sac is noted in the lower pelvis.
  - The lesion is continuous with the Posterolateral aspect of the urinary bladder.
  - Delayed imaging (approximately 30 minutes post-contrast) demonstrates filling of the diverticulum, which also shows complete or near-complete
  - emptying on post-void images.
  - No focal wall thickening or Intracavitary solid components are identified to suggest malignancy.

Fig. 5: Output of Agentic Workflow - Diagnostic Report

The diagnostic report is structured to provide a comprehensive overview of the patient's condition The report begins with the **primary diagnosis section** identifies the main condition, in this case, a herniated bladder diverticulum in the lower pelvis, supported by radiologic evaluation.

The **differential diagnoses** section lists potential alternative diagnoses, such as inguinal hernia with bladder involvement, urachal cyst, and pelvic cystic masses, with explanations for each. **Radiologic findings** summarize key imaging results, highlighting the characteristics o f the lesion and its implications.

Recommendations and next steps provide a plan for further imaging, urological consultation, laboratory evaluation, and patient monitoring. Management and treatment considerations discuss options like observation, surgical intervention, and symptomatic treatment based on the patient's condition. The summary of key insights offers a concise interpretation of imaging findings, differential considerations, and recommended next steps. Finally, the conclusion reinforces the diagnosis and suggests a collaborative management approach to optimize patient outcomes.

For evaluations, traditional NLP metrics such as ROUGE, BLEU, and even METEOR fail to capture the semantic nuances and context-dependent implications critical in clinical diagnostics. Their limitations highlight that standard evaluation methods are inadequate for our domain, prompting us to use DeepEval—an evaluation framework with metrics tailored specifically for assessing medical diagnostic accuracy and reliability.

Because medical errors can have serious consequences, we use higher thresholds, minimizing the risk of misleading or unsafe outputs. We selected the following metrics for our evaluation.

G-Eval: We employed this metric to simulate chain-of-thought reasoning, ensuring our outputs are both clinically coherent and accurate. Our model achieved a G-Eval score of 0.898.

Hallucination Detection: This metric identifies instances where the model generates unsupported or misleading content. Our model achieved a hallucination score of 0.667.

Faithfulness: This measure verifies that the output aligns closely with the reference clinical data. Our model achieved a perfect faithfulness score of 1.0. As observed in Table V, these metrics are crucial in the medical context, where even slight deviations can lead to significant diagnostic errors.

TABLE V: Performance Evaluation Metrics

Metric	Score	Threshold
GEval	0.898	0.85
Hallucination	0.667	0.90
Faithfulness	1.000	0.90

VI. Conclusion

In conclusion, our Vision-Language Model (VLM), fine-tuned with LoRA for medical image analysis, has shown notable gains in diagnostic accuracy, precise anatomical references, and streamlined resource usage. Quantitatively, our approach achieved an average GEval score of 0.75 and a faithfulness score of 0.79, demonstrating significant performance improvements over the base model. Building on these results, we introduced a multi-agent framework organized in a sequential chat model, where each specialized agent—ranging from data extraction to recommendation—processes thoroughly analyzed inputs before handing them off. This integrated system yielded exceptional results with a GEval score of 0.898 and perfect faithfulness at 1.000, exceeding our high clinical thresholds. By integrating the Vision Language Model's refined outputs with orchestrated multi-agent reasoning, the system produces cohesive, context-rich diagnostic reports tailored to each step in the clinical workflow. This integrated approach effectively enhances automated decision support in healthcare, showing substantial promise for improving clinical outcomes through both qualitative improvements and quantitatively validated performance metrics.

A. Future Work

In the future, we plan to extend our fine-tuned Vision-Language Model to cover a wider range of imaging modalities (e.g., MRI, PET, ultrasound) and larger, more diverse clinical populations. This broader validation will help establish real-world reliability and pave the way for integration into systems like Radiology Information Systems (RIS) or Electronic Health Records (EHR). Building on these enhancements, we also aim to expand our multi-agent architecture by exploring frameworks such as LangChain and CrewAI, allowing for an

even broader spectrum of medical scenarios and more complex decision-making workflows. Alongside more agents, we envision adding new tools—like dosage calculators, web search, and advanced visualization modules—to create a comprehensive environment that not only streamlines routine diagnoses but also manages high-complexity cases by routing them to specialized agents or human specialists when needed.

REFERENCES

[1] M. Ayaz, M. Khan, M. Saqib, A. Khelifi, M. Sajjad, and A. Elsaddik, "Medvlm: Medical vision-language model for consumer devices," *IEEE Consumer Electronics Magazine*, 2024.

[2] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. Seco de Herrera *et al.*, "Rocov2: Radiology objects in context version 2, an updated multimodal image dataset," *Scientific Data*, vol. 11, no. 1, p. 688, 2024.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[4] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu *et al.*, "Autogen: Enabling next-gen llm applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[5] J. Li, Y. Lai, W. Li, J. Ren, M. Zhang, X. Kang, S. Wang, P. Li, Y.-Q. Zhang, W. Ma *et al.*, "Agent hospital: A simulacrum of hospital with evolvable medical agents," *arXiv preprint arXiv:2405.02957*, 2024.

[6] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2022, 2022, p. 3876.

[7] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 353–367.

[8] R. Xu, W. Shi, J. Wang, J. Zhou, and C. Yang, "Medassist: Llm-empowered medical assistant for assisting the scrutinization and comprehension of electronic health records," 2025.

[9] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "Radbert: adapting transformer-based language models to radiology," *Radiology: Artificial Intelligence*, vol. 4, no. 4, p. e210258, 2022.

[10] J. Sayyad Shirabad, S. Wilk, W. Michalowski, and K. Farion, "Implementing an integrative multi-agent clinical decision support system with open source software," *Journal of medical systems*, vol. 36, pp. 123–137, 2012.

[11] Y. Bazi, M. M. A. Rahhal, L. Bashmal, and M. Zuair, "Vision–language model for visual question answering in medical imagery," *Bioengineering*, vol. 10, no. 3, p. 380, 2023.

Lang

by Vinutha N

Submission date: 16-Mar-2025 09:13PM (UTC+0530)
Submission ID: 2615993953
File name: IEEE_Connect_2025.pdf (484.9K)
Word count: 3278
Character count: 19921

Enhancing Medical Diagnostics with Vision-Language Models and Agents

1st Sri Bharath Sharma P
CSE (AI&ML)
Dayananda Sagar University
Bangalore, India
bharathperala28@gmail.com

2nd Ratan Ravichandran
CSE (AI&ML)
Dayananda Sagar University
Bangalore, India
ratan152004@gmail.com

3rd Shriyans Shriniwas Arkal
CSE (AI&ML)
Dayananda Sagar University
Bangalore, India
shriyans.arkal07@gmail.com

4th Sayli Pankaj Bande
CSE (AI&ML)
Dayananda Sagar University
Bangalore, India
gaylibande19@gmail.com

5th Yudhajit Jana
CSE (AI&ML)
Dayananda Sagar University
Bangalore, India
yudhajit.j12@gmail.com

6th Vinutha N
CSE (AI&ML)
Dayananda Sagar University
Bengaluru, India
vinutha.n-aiml@dsu.edu.in

***Abstract*—Commercial vision models work well for general imaging tasks but struggle with radiographic diagnosis because they miss the domain-specific details needed for accuracy. This is a major issue since radiography is essential for revealing complex anatomical structures. This paper addresses these limitations by proposing a novel multimodal approach that adapts the Llama 3.2-11B-Vision model using Low-Rank Adaptation (LoRA) on the ROCOv2: Radiology Objects in COntext Version 2 dataset. Furthermore, to overcome the diagnostic shortcomings of standalone vision-language systems, we introduce an end-to-end multi-agent framework that ingests the diagnostic output of the fine-tuned vision-language model, then uses chain-of-thought reasoning and sequential processing to perform specialized tasks such as patient history extraction, case data interpretation, lab report analysis, summarization, and diagnosis. This integrated approach enhances overall diagnostic performance, bridging the gap between general-purpose vision models and the specialized requirements of accurate radiographic diagnosis.**

***Index Terms*—Vision models, Radiographic diagnosis, Low-Rank Adaptation (LoRA), Chain-of-thought reasoning**

I. Introduction

Radiology is pivotal for diagnosis and treatment planning, yet increasing data complexity—from high-resolution CT, MRI, and X-ray images to comprehensive clinical cases—has exposed significant shortcomings in traditional diagnostic workflows. Existing systems rely on single-agent or monomodal approaches that inadequately integrate visual and textual information, leading to suboptimal diagnostic accuracy and efficiency. Our project tackles these challenges by fine-tuning a state-of-the-art vision-language model using techniques such as Low-Rank Adaptation (LoRA) and embedding it within a multi-agent AI framework. This framework leverages specialized agents—each designed to handle discrete tasks like patient history extraction, lab data interpretation, and imaging correlation—allowing for chain-of-thought reasoning and cross-validation that enhances both interpretability and robustness. Unlike conventional models that process diverse inputs in a single flow, our multi-agent system distributes the workload among task-specific agents, thereby reducing oversimplification and error propagation commonly seen in current approaches. This targeted division of labor not only improves diagnostic accuracy in complex clinical scenarios but also offers technical advantages such as efficient parameter tuning, dynamic adaptation to new data distributions, and enhanced performance in outlier cases. Ultimately, our approach demonstrates clear improvements over existing methods, paving the way for more reliable and scalable clinical applications.

II. Literature Review

Ayaz et al. [1] introduce MedVLM, a vision-language model integrating Florence-2 and LLaMA-2 via LoRA, optimized for medical applications such as Visual Question Answering (VQA) and medical report generation. Evaluations on Rad-VQA show superior accuracy over specialized and generalist models, with radiologists validating 74% of the generated reports as high quality.

Rückert et al. [2] present ROCOv2 (Radiology Objects in COntext Version 2), a multimodal dataset comprising of 79,789 radiological images with curated captions and medical concepts. It adds 35,705 new images over the previous dataset (ROCO) published in 2018 and supports training image annotation models based on image-caption pairs, or for multi-label image classification.

Lang

ORIGINALITY REPORT

16%	13%	12%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	4%
2	www.fh-dortmund.de Internet Source	1%
3	Submitted to University of Warwick Student Paper	1%
4	www.nature.com Internet Source	1%
5	assets-eu.researchsquare.com Internet Source	1%
6	Carmen De Maio, Giuseppe Fenza, Domenico Furno, Teodoro Grauso, Vincenzo Loia. "Privacy-Preserving Healthcare Data Interactions: A Multi-Agent Approach Using LLMs", Journal of Communications Software and Systems, 2025 Publication	1%
7	Submitted to University of Edinburgh Student Paper	1%
8	Submitted to University of Pretoria Student Paper	1%
9	Zhang, Jiyin. "Knowledge-Infused LLM Application in Data Analytics: Using Mindat as an Example", University of Idaho, 2025 Publication	<1%
10	Submitted to Dayananda Sagar University, Bangalore Student Paper	<1%

11	S N Siri, H B Divyashree, S Pushpa Mala. "The Memorable Assistant: An IoT-Based Smart Wearable Alzheimer's Assisting Device", 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021 Publication	<1%
12	Submitted to University of Glasgow Student Paper	<1%
13	www.philschmid.de Internet Source	<1%
14	www.arxiv-vanity.com Internet Source	<1%
15	www.springerprofessional.de Internet Source	<1%
16	H.L. Gururaj, Francesco Flammini, J. Shreyas. "Data Science & Exploration in Artificial Intelligence", CRC Press, 2025 Publication	<1%
17	interviewprep.org Internet Source	<1%
18	Submitted to AUT University Student Paper	<1%
19	Xiong, Zinan. "Advancing Healthcare Through Deep Learning: From Disease Recognition to Human Pose Estimation in Imaging and Video Data.", University of Massachusetts Lowell Publication	<1%
20	ceur-ws.org Internet Source	<1%
21	semantic-web-journal.net Internet Source	<1%
22	Tsz Kin Chan, Ngoc-Duy Dinh. "ENTAgents: AI Agents for Complex Knowledge	<1%