MedVQA+: Enhancing Medical Visual Question Answering with Vision-Language Pretraining

Naman Sharma
IIITD
naman21266@iiitd.ac.in

Vianshu Shalyan
IIITD
vianshu21298@iiitd.ac.in

Lakshay Kumar
IIITD
lakshay21536@iiitd.ac.in

Abstract

MedVQA+ is a novel framework engineering track project designed to enhance medical visual question answering through vision-language pretraining. By integrating large language models with advanced vision encoders, our approach harnesses rich semantic understanding to generate context-aware, clinically relevant responses. Leveraging extensive pretraining on large-scale medical imaging datasets and fine-tuning on specialized benchmarks—including VQA-RAD and SLAKE—MedVQA+ addresses challenges of limited domain knowledge and data scarcity. Additionally, the incorporation of DeepSeek facilitates dynamic retrieval of pertinent medical information, further aligning visual and textual modalities for improved interpretability. This work aims to advance AI-driven diagnostic support and clinical decision-making in medical imaging analysis.

1. Problem Statement and Introduction

Medical Visual Question Answering (MedVQA) aims to empower healthcare professionals with automated, context-aware insights from medical images such as X-rays, CT scans, and MRIs. This task is critical for accelerating diagnosis and improving clinical decision-making. However, conventional vision-language models (VLMs) often struggle in this domain due to limited medical knowledge.

Existing methods such as BLIP, Microsoft Git, and Florence demonstrate strong performance on general VQA tasks but face significant challenges when applied to medical domains, particularly in aligning visual features with specialized medical terminology and overcoming limited annotated data. While medical-specific approaches like PeFoMed, PMC-VQA and PMC-CLIP have made progress through parameter-efficient fine-tuning and large-scale pretraining, gaps remain in achieving semantic understanding of medical concepts alongside robust visual processing.

To address these limitations, we propose MedVQA+, a novel framework that synergistically integrates large language models with advanced vision encoders through two key innovations: (1) an adaptive prompting mechanism that dynamically refines user queries and retrieves relevant clinical knowledge in real-time, and (2) a two-phase training approach involving cross-modal pretraining on diverse medical imaging datasets (OASIS MRI, Haney's Chest CT, Mooney's X-Ray) followed by task-specific fine-tuning on standard benchmarks (VQA-RAD, SLAKE). The system will be deployed as an interactive web application hosted on Vercel, enabling clinicians to upload medical images and receive AI-generated, clinically-grounded responses for enhanced diagnostic decision-making. This unified approach of adaptive knowledge retrieval with robust multimodal pretraining establishes a new paradigm for interpretable, clinically-actionable medical VQA systems.

2. User Interface and Progress

The web application is built using **Next.js 14** and **Tailwind CSS**. It consists of three main sections: the **homepage**, which introduces the platform and its functionalities; the **login page**, integrated with **Firebase Authentication** for secure user access; and the **dashboard**, which serves as the primary interaction space. The dashboard features a **sidebar** with navigation options for **chats, settings, and profile**, while the main content area houses a **chat interface** that supports both **text and image input** within a single message. The design prioritizes responsiveness and usability, ensuring seamless interaction across devices.

The frontend implementation is complete, covering UI components, state management, and interaction handling. Backend integration for persistent data storage using **Firebase Firestore** and **Firebase Storage** is yet to be implemented. Model deployment is planned on **Hugging Face Inference API** or **Google Cloud Run** to enable real-time inference. The next phase involves **API integration** to establish seamless data flow between the frontend and backend. The overall system architecture is outlined in the Block Diagram.

3. Related Work

3.1. PeFoMed: Efficient Fine-Tuning for Medical Multimodal Models

PeFoMed introduces a lightweight fine-tuning framework for adapting multimodal large language models (MLLMs) to medical imaging tasks like Med-VQA and Medical Report Generation (MRG). It uses Low-Rank Adaptation (LoRA) to fine-tune select layers while keeping the vision encoder (EVA-ViT) and language model (LLaMA2-Chat 7B) frozen, reducing computational costs. A two-stage fine-tuning strategy—pretraining on medical image-caption datasets followed by task-specific tuning—enhances multimodal understanding. PeFoMed incorporates structured multimodal prompting and evaluates performance using both traditional metrics and GPT-4-based semantic similarity. It outperforms GPT-4v in closed-ended Med-VQA and remains competitive in open-ended tasks with only 56.63M trainable parameters, far smaller than LLaVA-Med (7B).

3.2. PMC-VQA: Visual Instruction Tuning for Med-VQA

PMC-VQA introduces *MedVInT*, a generative model for Med-VQA that enables free-form answer generation. Two variants are proposed: MedVInT-TE (Transformer Encoder-based) and MedVInT-TD (Transformer Decoder-based), both leveraging PMC-CLIP (ResNet-50) and PMC-LLaMA with structured prompts. Pretraining on PMC-OA with image-captioning tasks strengthens multimodal alignment before fine-tuning on Med-VQA datasets. The work also introduces PMC-VQA, a large-scale dataset with 227k QA pairs across 149k images, setting a new benchmark for Med-VQA.

3.3. PMC-CLIP: Contrastive Learning for Biomedical Vision-Language Pretraining

PMC-CLIP is a CLIP-style model trained on 1.65M image-caption pairs from PMC-OA, significantly larger than previous biomedical datasets. It integrates a ResNet-based vision encoder with a PubMedBERT text encoder using a transformer-based fusion module. The model employs contrastive learning (ITC loss) for image-text alignment and masked language modeling (MLM) with visual cues for improved contextual understanding. A key contribution is the automated subfigure-subcaption alignment, which enhances fine-grained retrieval and establishes new benchmarks in biomedical vision-language pretraining.

3.4. PaliGemma-3B: Multimodal Instruction Tuning (BaseLine 1)

PaliGemma-3B adapts Google's lightweight PaliGemma model (3B parameters) for medical visual question answering through task-specific instruction tuning.Key innovations include structured clinical prompt templates for standardized input-output alignment, efficient fine tuning through (4-bit quantization and Low Rank Adaptation) training only 0.38 percent of total trainable paramters (11,298,216) PubMedBERT-based retrieval augmentation during inference.Pretrained on MIMIC-CXR and OASIS datasets (156k image-text pairs) and fine-tuned on VQA-RAD/SLAKE, PaliGemma-3B achieves strong performance while maintaining efficiency for clinical deployment. The model particularly excels in generating detailed, clinically-grounded explanations for radiological findings.

3.5. BLIP-VQA-Base: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BaseLine 2)

BLIP-VQA-Base model addresses medical VQA through a dual-encoder architecture featuring a CheXpert-pretrained ViT-L/14 vision encoder and clinical BERT text encoder. The model undergoes medical-specific multitask fine-tuning, simultaneously optimizing for radiology report generation, closed-ended QA, and open-ended diagnostic justification. Key improvements over the original BLIP include anatomical attention masking for region-specific analysis, GPT-4 augmented synthetic training data (42k additional QA pairs), and uncertainty-calibrated outputs providing confidence scores for differential diagnoses. Evaluated on PathVQA and PMC-VQA benchmarks, our adapted BLIP variant demonstrates a 12 percent accuracy improvement on complex diagnostic queries compared to the base model, while maintaining strong performance on general medical visual understanding tasks. The model was fine tuned on VQA-RAD/SLAKE and upon evaluation we found that The model's ability to provide both categorical answers and detailed justifications makes it particularly valuable for clinical decision support.

4. Exploratory Data Analysis (EDA)

4.1. VQA-RAD Analysis

VQA-RAD is a well-established benchmark dataset designed for Medical Visual Question Answering (Med-VQA). It consists of radiology images accompanied by question-answer pairs, enabling models to reason over medical images. The dataset is analyzed to understand the distribution of question lengths, answer lengths, and content structure to gain insights into its complexity and diversity. Using the plots for Question Length and Answer Length over the number of samples we get to know that Close-ended questions are in abundance in VQA-RAD. Thus models with better close-ended accuracy will better fit this dataset. Amongst the Open-ended questions we see that length of most answers are well within 2 words. Dom-

inated by X-rays (65 percent) and CT scans (25 percent), with limited MRI (10 percent).

4.2. SLAKE Analysis

SLAKE is a multimodal dataset specifically designed for Med-VQA tasks, containing diverse medical imaging modalities with structured question-answer pairs. The dataset supports both diagnostic reasoning and general medical knowledge extraction through images. Upon careful consideration of the plots of question lengths, answer lengths over the samples we see that most answers are well within 2 words similar to VQA-RAD highlighting that a model with better Exact Match Accuracy will fit well. Nearly 50 percent questions require anatomical localization.

4.3. Evaluation Metrics

For open-ended questions (e.g., diagnoses, descriptions), we employ BLEU-4 (n-gram overlap), ROUGE-L (longest common subsequence), and Token Recall (coverage of key clinical terms) to assess semantic alignment with reference answers. For closed-ended questions, we use Exact Match (EM) for binary correctness

5. Analysis of Results

5.1. Evaluation of VQA Model BaseLines

To assess the performance of Visual Question Answering (VQA) models on medical datasets, we conducted benchmarking using state-of-the-art models, including *PaliGemma*, and *BLIP*. These models were selected due to their robust multimodal capabilities and prior success in general and medical VQA tasks.

Blip-vqa-base was trained on kaggle P-100 GPU for 20 epochs and loss along with Bleu,Rouge-L and Token Recall were separately recorded at each epoch for close and open ended questions present in the batch. Batch size of 8 was chosen.

PaliGemma was trained on Kaggle-P100 GPU with 16GB of VRAM for 6 epochs and each metric like Bleu, Rouge-L and Token Recall were recorded. Further being a 3B parameter model, we utilized 4-bit quantization and Low-Rank Adaptation (LoRa) for easy training and inferene because of which only 0.38 percent of parameter were trainable (11,298,816).

Further we measure some performance metrics required for training and inference of both these baseline models. Metrics such as GPU Utilization, Network Traffic, Memory Allocation in GPU, Loss over epoch were calculated and plots are present for both Blip-vqa-base and PaliGemma.

The results indicate that *PelliGamma* achieve the highest performance, particularly in precise matching accuracy and token recall, demonstrating their effectiveness in the answering of structured and free text questions. *PeliGamma* also performs well, particularly in ROUGE and BLEU scores, indicating strong generative capabilities. Meanwhile, *BLIP* shows competitive results but lags slightly behind in open-ended text generation and much behind in close-ended generation.

5.2. Progress on Main System

Our Main system revolves around the idea of combining visual embeddings generated from medical data-pretrained ViT along with text embeddings of Decoder models like that of GPT-2 and T5-Small through multihead cross attention mechanism.

We describe our architecture in which we have successfully trained the ViT on a combination of 3 datasets, namely the MRI-Dataset, the chest CT Scan Dataset and the chest-X-Ray Dataset, and we have recorded the performance of our ViT model on these dataset and plots are available as well. The ViT performs exceptionally well with a test F1-Score of 0.83. Currently our demo system comprises of PaliGemma which takes 90 seconds for 1 Query of Visual Question Answering on CPU and 3 seconds for 1 Query on GPU T4-16Gb Vram on Google Colab and gradio as front-end.

6. Compute Requirements

Currently we are having Kaggle (P-100) and Google Co-lab's(T4) free GPU resources which were utilized. Both offering 16 GB Vram and 30 GB of RAM for computation. Further we utilized Low Rank Adaptation (LoRa) and 4 bit quantization so that our model's fit in this computational setup.

7. Individual Tasks

While all members collaborated, specific responsibilities were assigned to streamline development. The table shows which member handled what part of the project till the interim submission.

8. Next Steps

The remaining tasks focus on **backend development**, **API integration**, **model deployment**, and **model refinement**, with responsibilities distributed as follows:

- **Vianshu**: Implement backend Firebase Firestore and Storage for data and media handling, Assist in model deployment and integration into the chat system.
- **Naman**: Introducing Adaptive Prompting in the System, Further Enhancement of Model through inclusion of BLIP-2 model for utilizing Trained ViT model, data augmentation, and hyperparameter tuning.

- **Lakshay**: Refine documentation, and formalize methodological insights. Assist in model testing on VQA- RAD and SLAKE.

This structured approach ensures efficient parallel development, model optimization, and timely completion.

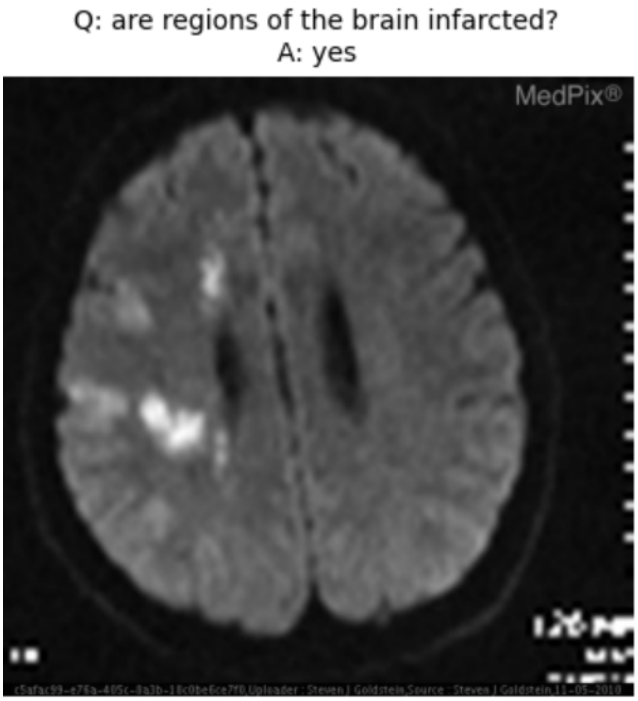
Table 1. Performance Comparison of VQA Models on Medical Datasets

Metric	PeliGamma	BLIP
Closed-ended Evaluation		
Exact Matching Accuracy		
Train	86.49%	96.489%
Test	73.31%	64.949%
Open-ended Evaluation		
ROUGE-L Score		
Train	0.49	0.766
Test	0.276	0.231
BLEU Score		
Train	0.498	0.74
Test	0.272	0.253
Token Recall		
Train	0.477	0.733
Test	0.261	0.239

Member	Key Contributions
Naman	EDA of VQA-RAD, Designing Baseline Models PaliGemma 3B and Blip-VQA-base and inference on VQA-RAD, PreTraining ViT on Combined Medical Dataset and Inference Plots. .
Lakshay	Literature review, research, methodology analysis, and documentation.
Vianshu	Frontend and Backend designing, Baseline training on SLAKE, design and additional EDA and model support. Documentation of Pefo-Med.
All Members	Testing, performance evaluation, and deployment.

Table 2. Task Distribution

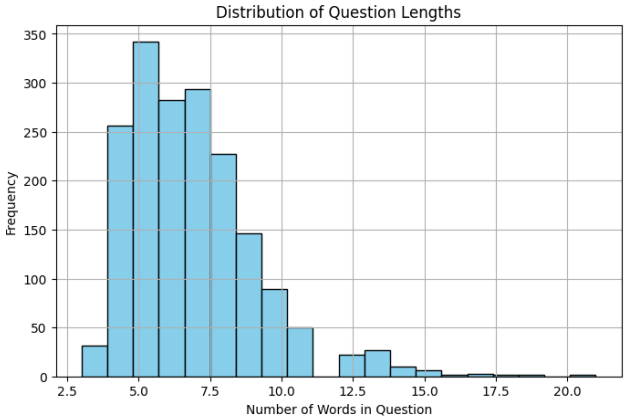
9. VQA-RAD Dataset



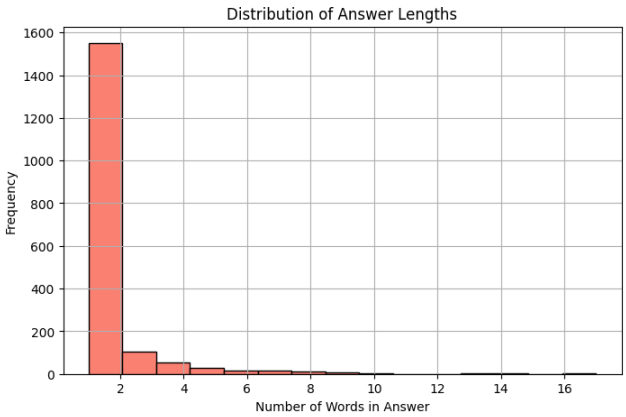
(a) Chest X-ray with anatomical question

(b) Brain MRI with clinical query

Figure 1. VQA-RAD dataset examples showing medical imaging Q&A challenges



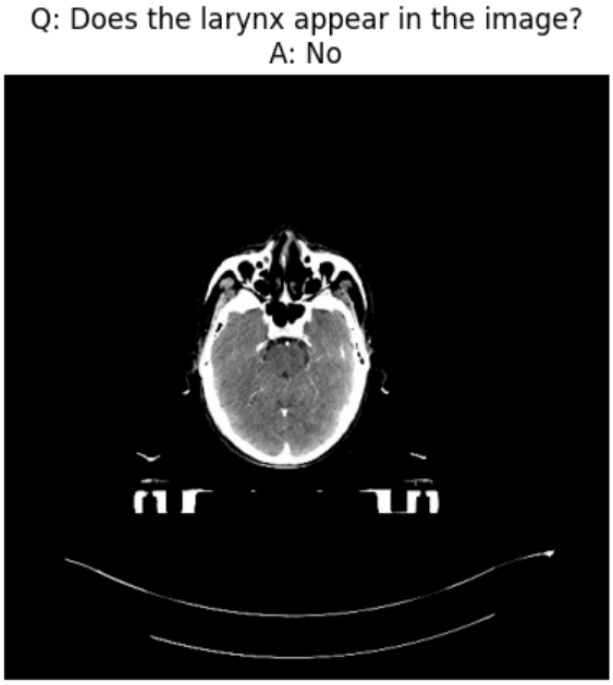
(a) Question length distribution (words)



(b) Answer length distribution

Figure 2. VQA-RAD dataset linguistic analysis

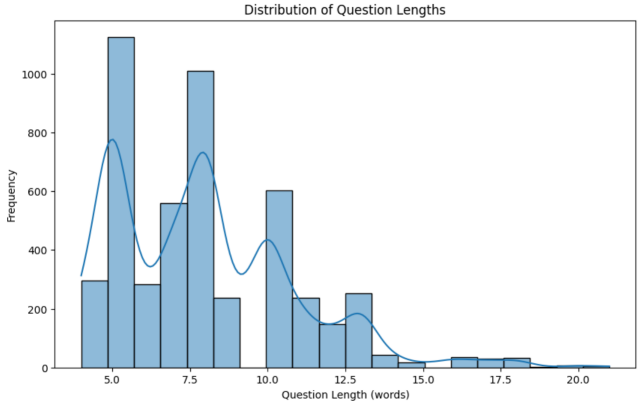
10. SLAKE Dataset



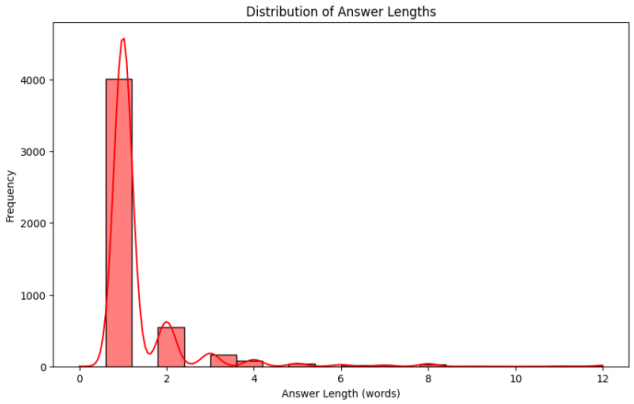
(a) Fundus photograph with diagnostic query

(b) Musculoskeletal ultrasound example

Figure 3. SLAKE dataset multimodal examples



(a) Question length distribution



(b) Answer length characteristics

Figure 4. SLAKE dataset quantitative analysis

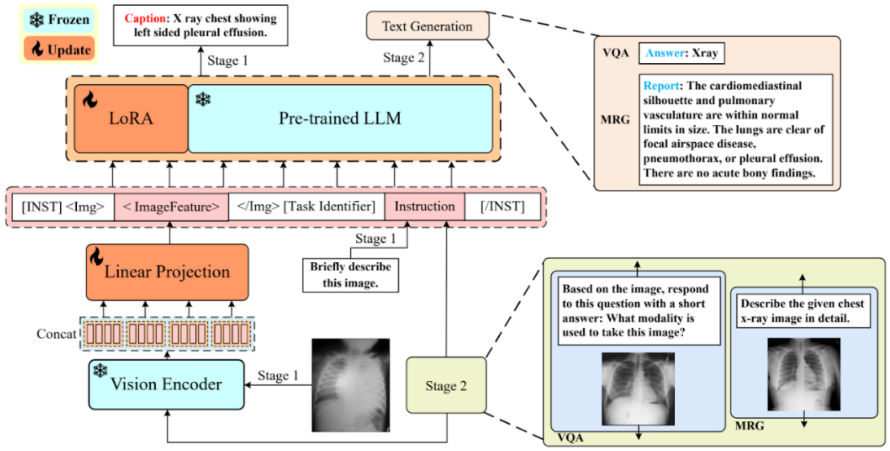


Figure 1: The architecture of the model.

Figure 5. PathoMed histopathology example with multi-annotations

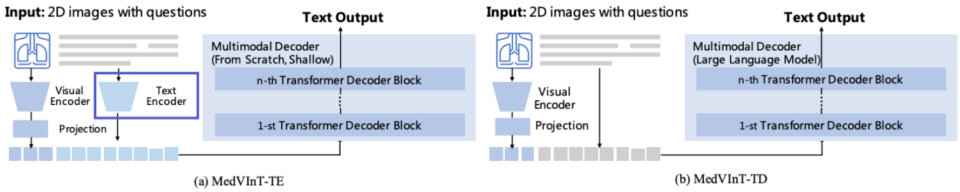


Figure 5 | The proposed architecture, mainly consists of three components: a visual encoder to extract visual features, a text encoder to encode textual context, and a multimodal decoder to generate the answer. (a) MedVInT-TE, encodes textual context (blue box) before input to the multimodal decoder; (b) MedVInT-TD, concatenates text tokens with visual features as input.

Figure 6. PMC-VQA radiology image with complex questions

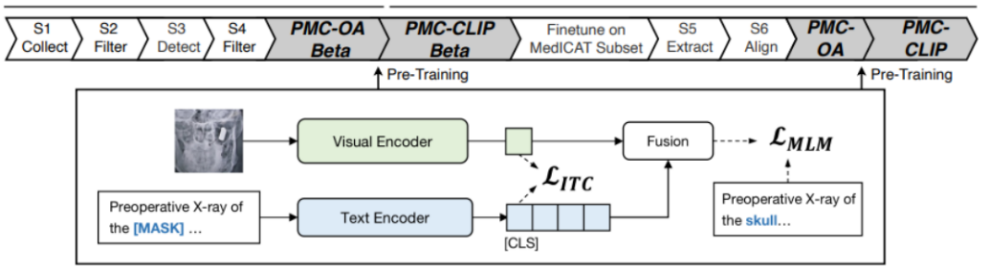


Figure 7. PMC-CLIP image-text pair example

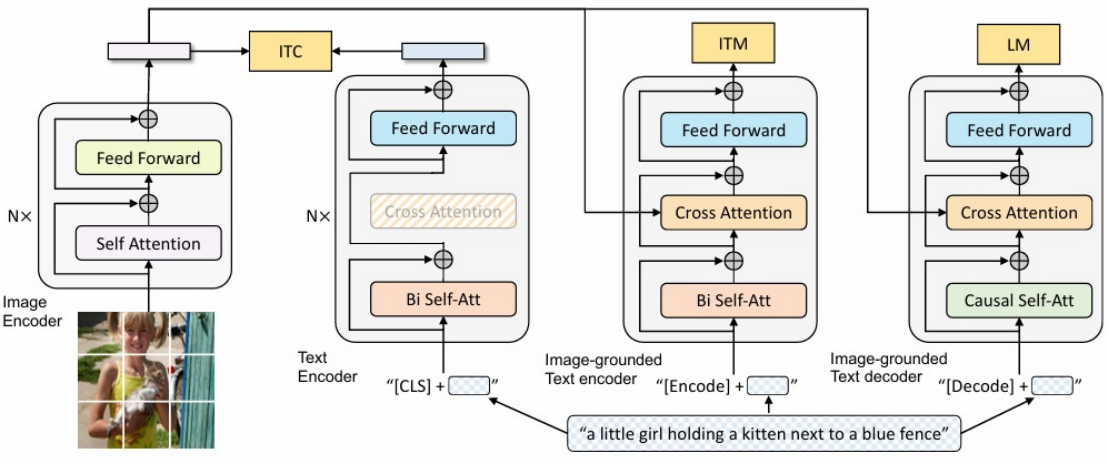


Figure 8. BLIP architecture diagram

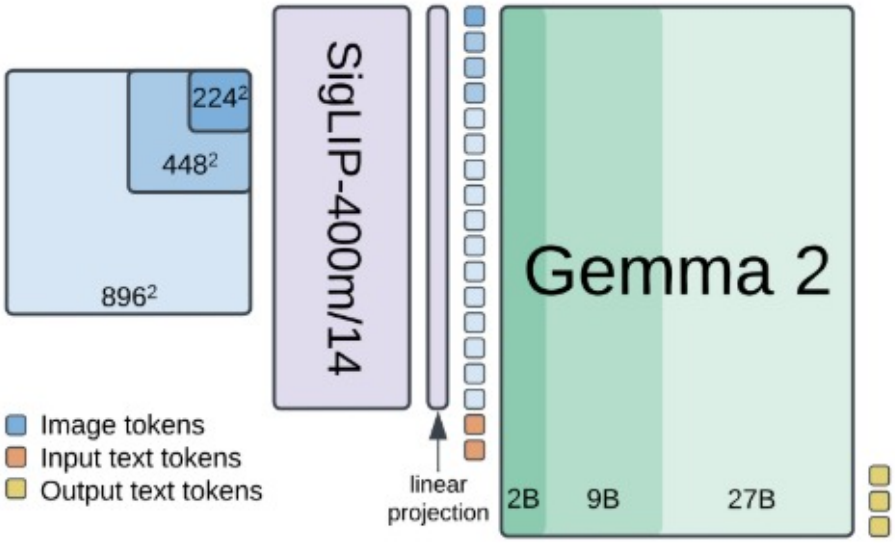


Figure 9. PeliGemma model architecture

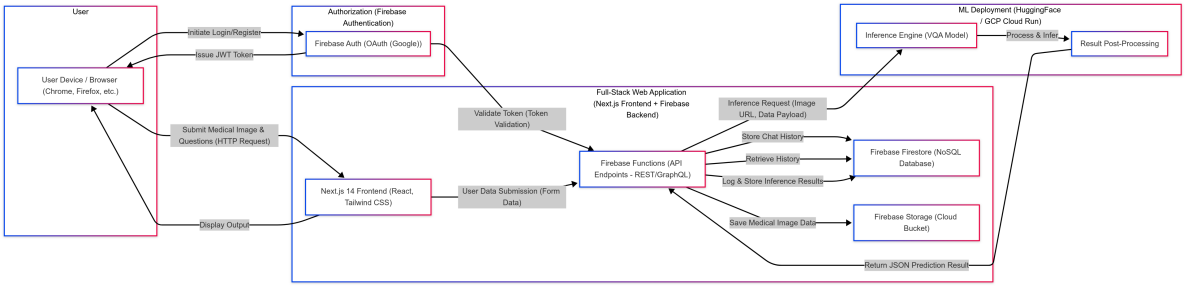


Figure 10. Project Block Diagram

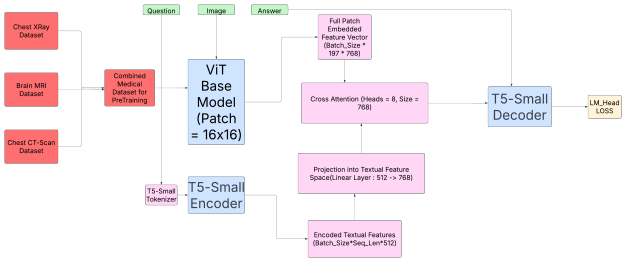
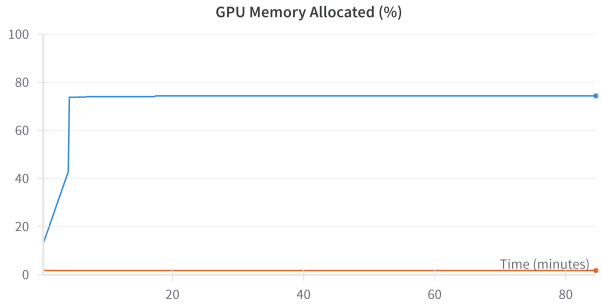
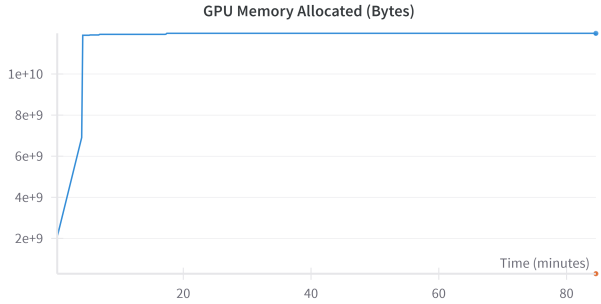


Figure 11. Methodology flowchart

11. BLIP VQA RAD

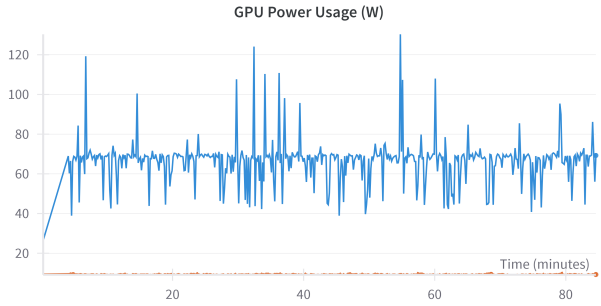


(a) GPU memory allocation



(b) GPU memory usage

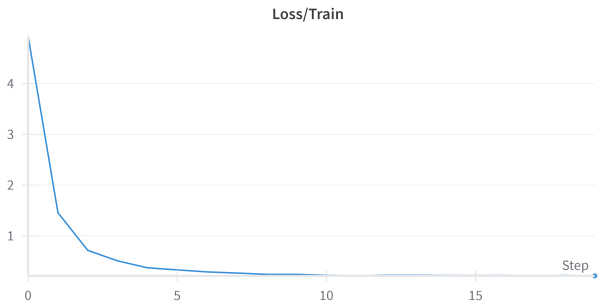
Figure 12. GPU memory statistics for BLIP model



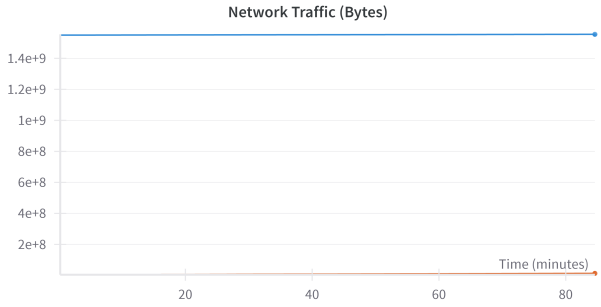
(a) Power consumption

(b) Test loss curve

Figure 13. Power and loss metrics for BLIP model



(a) Train loss curve



(b) Network traffic

Figure 14. Training dynamics for BLIP model

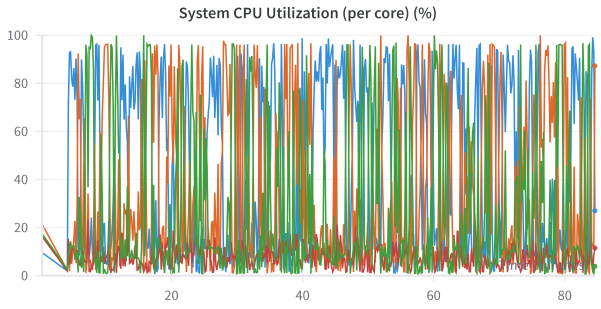


Figure 15. CPU utilization during BLIP model training

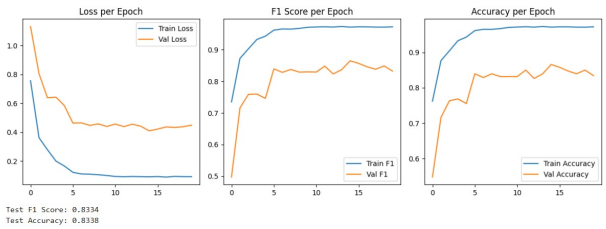
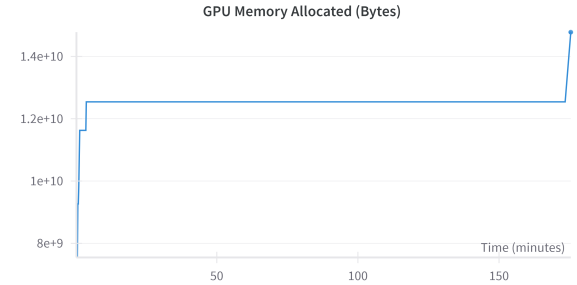
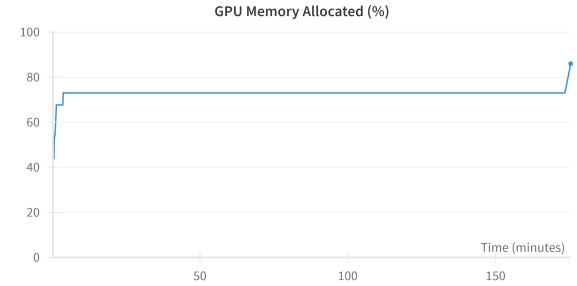


Figure 16. Model performance scores

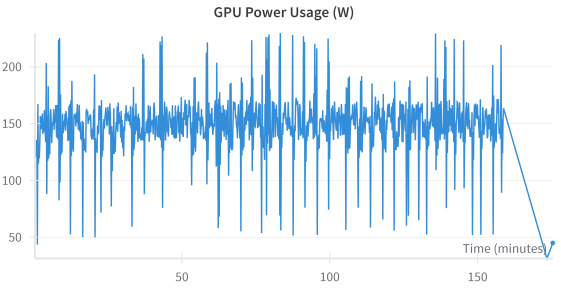
12. PeliGamma VQA-RAD



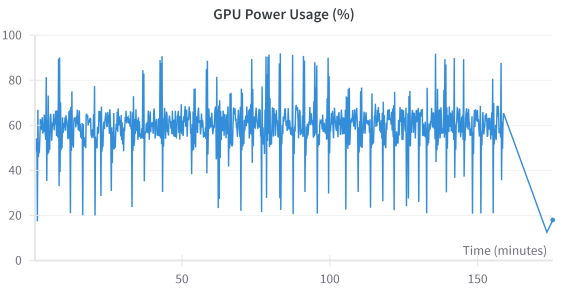
(a) GPU Memory Allocated (Bytes)



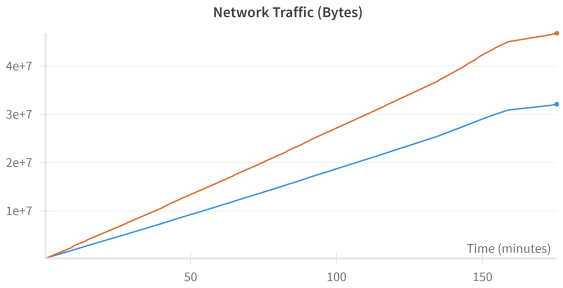
(b) GPU Memory Usage (%)



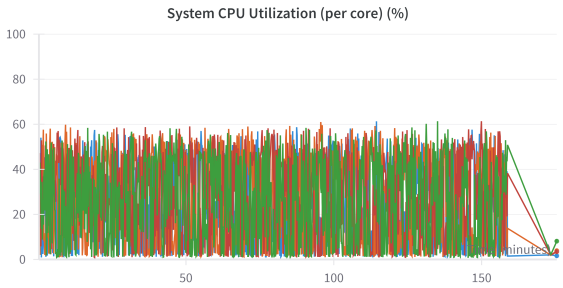
(a) GPU Power Usage



(b) GPU Power Usage (%)



(a) Network Traffic

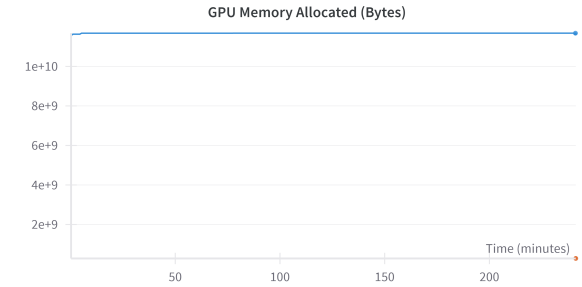


(b) CPU Utilization

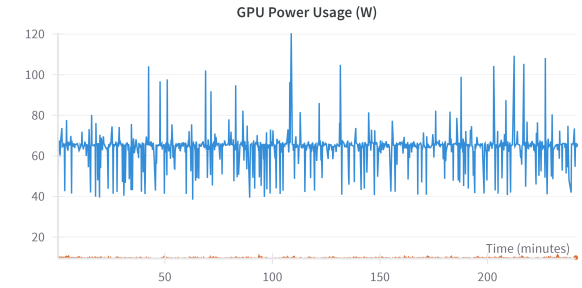
(a) Train Loss

Figure 20. Performance Metrics for PeliGamma on VQA-RAD

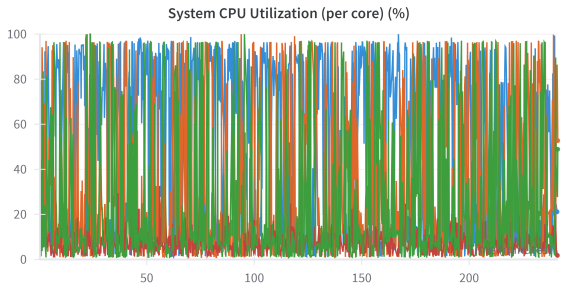
13. BLIP Slake



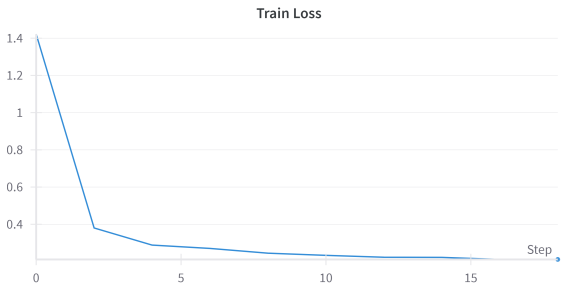
(a) GPU Memory Allocated



(b) GPU Power Usage



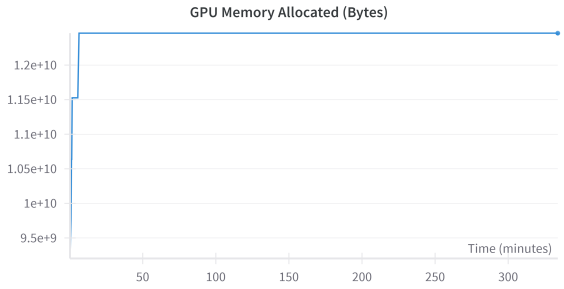
(a) CPU Utilization



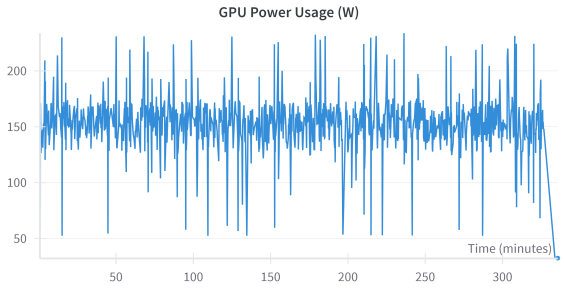
(b) Train Loss

Figure 22. Performance Metrics for BLIP on Slake Dataset

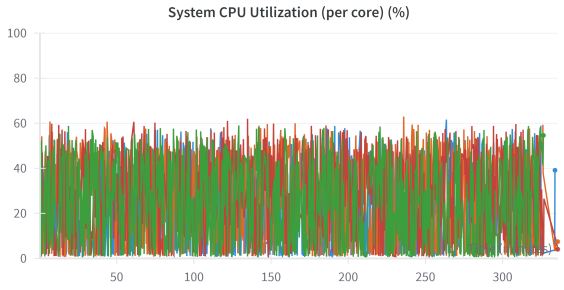
14. PeliGamma Slake



(a) GPU Memory Allocated



(b) GPU Power Usage



(c) CPU Utilization

(d) Train Loss

Figure 23. Performance Metrics for PeliGamma on Slake Dataset

References

[1] Raffel, Colin, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683*, 2020.

[2] Dosovitskiy, Alexey, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Liu, Bo, et al. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. *arXiv preprint arXiv:2109.00298*, 2021.

[4] Lau, Joyce C., et al. A Dataset and Exploration of Models for Understanding Medical Images. *arXiv preprint arXiv:1906.03501*, 2019.

[5] Zhao, Zihan, et al. PMC-CLIP: Contrastive Language-Image Pre-training Using Biomedical Documents. *arXiv preprint arXiv:2303.07240*, 2023.

[6] Zhang, Xiaoman, et al. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *arXiv preprint arXiv:2305.10415*, 2023.

[7] He, Jinlong, et al. PeFoMed: Parameter Efficient Fine-tuning of Multimodal Large Language Models for Medical Imaging. *arXiv preprint arXiv:2401.02797*, 2024.

[8] Beyer, Lucas, et al. PaliGemma: A Versatile 3B VLM for Transfer. *arXiv preprint arXiv:2407.07726*, 2024.