Research Plan: Multi-Image Medical VQA Improvement

Step-by-Step Research Plan

Step 1: Dataset Acquisition

What: Download MedFrameQA benchmark dataset
Source: https://huggingface.co/datasets/SuhaoYu1020/MedFrameQA
Content: 2,851 multi-image medical questions with 2-5 images each
Coverage: 9 body systems, 43 organs, multiple modalities (CT, MRI, X-ray, Ultrasound)

Step 2: Test 6 State-of-the-Art Medical VQA Models

Model 1: LLaVA-Med v1.5 (Microsoft)

- **Download**: https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b
- **Why**: Most popular medical VQA model (7B parameters)

Model 2: BiomedCLIP + LLaMA-3

- **Download**: https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224
- **Why**: Best medical image encoder (8B parameters)

Model 3: MedGemma-4B (Google)

- **Download**: https://huggingface.co/google/medgemma-4b-it
- **Why**: Efficient 4B-parameter model

Model 4: Bio-Medical-LLaMA-3-8B

- **Download**: https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B
- **Why**: Specialized LLaMA-3 with medical fine-tuning

Model 5: Qwen2.5-VL-7B Medical

- **Download**: https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
- **Why**: Latest general vision-language capabilities

Model 6: PMC-VQA Model (Research Implementation)

- **Source**: Implementation from PMC-VQA paper (2024)

Step 3: Systematic Testing Protocol

3.1 Test Each Model on MedFrameQA

```
For each model:
  - Load model and MedFrameQA dataset
  - Test on initial 200 questions, then full set
  - Measure overall accuracy
  - Track performance by:
    * Number of images (2-5)
    * Body system
    * Modality
    * Question type
```

3.2 Expected Results

- **All models will show <55% accuracy**
- **20-30% drop** vs single-image tasks
- **Weaknesses**: cross-image reasoning

Step 4: Detailed Failure Analysis

4.1 Categorize Failure Types

Type 1: Cross-Image Attention Failure

- Ignores additional frames

Type 2: Evidence Aggregation Failure

- Cannot combine findings across frames

Type 3: Temporal Reasoning Failure

- Fails to interpret progression

Type 4: Spatial Relationship Failure

- Misses anatomical connections

Type 5: Error Propagation

- Early mistake cascades to final answer

4.2 Pattern Analysis

```
failure_analysis = {
    "most_common_failure_type": "",
    "hardest_body_system": "",
    "hardest_modality": "",
    "performance_by_image_count": {}
}
```

Step 5: Decision Point - Confirm Problem Exists

If all models show <55% accuracy:

- Confirm universal gap
- Proceed to solution development

If some models >60% accuracy:

- Problem may be model-specific
- Pivot focus or refine analysis

Step 6: Problem Statement Finalization

Problem Statement: "State-of-the-art medical VQA models fail at reliable multi-image reasoning, achieving <50% accuracy on MedFrameQA despite strong single-image performance, hindering clinical deployment."

Research Objective: "Investigate and develop innovative approaches—including but not limited to fine-tuning, architectural modifications, attention mechanisms, or other solutions—to significantly improve multi-image medical reasoning performance while preserving existing single-image capabilities."

Step 7: Solution Development Strategy

7.1 Literature Review

- Explore multi-image attention, temporal reasoning, domain adaptation, training strategies

7.2 Solution Exploration

- **Cross-Image Attention Enhancement**
- **Sequential Reasoning Modules**
- **Clinical Evidence Fusion**
- **Error-Resistant Architectures**
- **Novel Training Strategies**

7.3 Implementation

- Build prototype modules

- Fine-tune base models on targeted failure modes
- Evaluate improvements across all models

Next Steps

1. Begin Step 1: Download dataset and set up environment
2. Execute Step 2 for initial validation
3. Review results and decide on Step 5 outcome

This clear, concise plan focuses on immediate validation followed by systematic solution development.