Fine-Grained Multimodal RAG: Enhancing Retrieval with Object-Level Representations

CS6101 Project 7 — Manaswini Talagadadivi, Shen Ting Ang, Huang Chao Ming, Benjamin Goh, Estelle Sim

1. ABSTRACT

Large Vision-Language Models (LVLMs) struggle to effectively utilize retrieved visual knowledge for complex reasoning tasks, particularly those involving changes in perspective, scope, or occlusion, as demonstrated by the MRAG-Bench. Our llava-onevision-7b baseline model achieved a strong initial accuracy of 58.2%.

We introduce an **Object Detection Enhancement to the MRAG-Bench evaluation** pipeline to push performance beyond the existing baseline by providing explicit visual grounding. This utilizes the **DETR (DEtection TRansformer) mode**l to analyze images and convert visual content (objects, counts, and spatial layout) into **structured text descriptions.** This **structured analysis** is used to create an **Enhanced Prompt** that guides the LLaVA model's reasoning.

2. INTRODUCTION

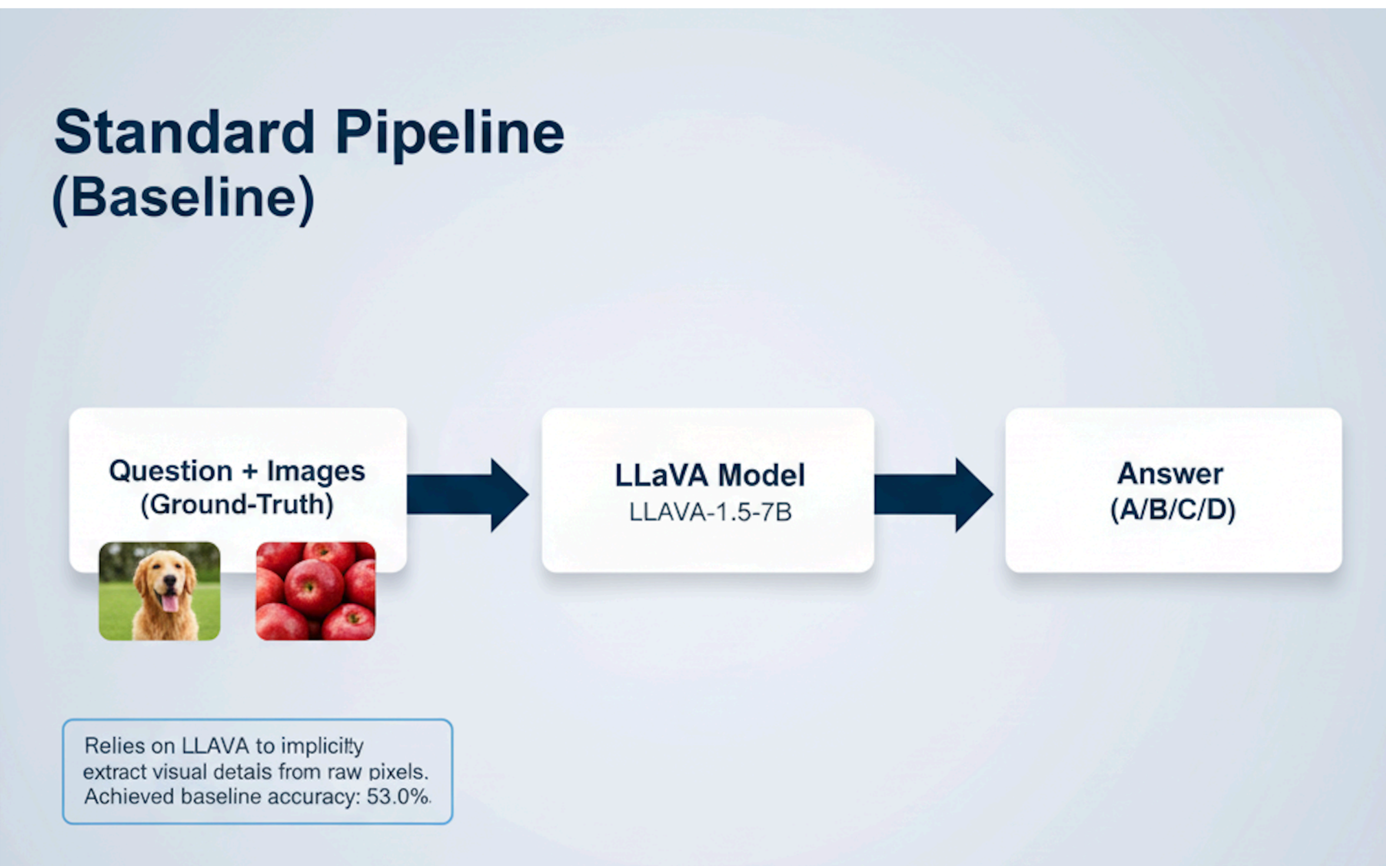
MRAG-Bench Context: The MRAG-Bench is a vision-centric evaluation benchmark designed to test LVLMs' ability to utilize visually augmented knowledge, consisting of 1,353 multiple-choice questions across 9 distinct scenarios. Key challenge scenarios include *Angle*, *Partial View*, and *Occlusion*.

Achieved Baseline: Our llava-onevision-7b implementation has achieved a measured baseline accuracy of 58.2%, successfully meeting the lower boundary of the 53–59% target range.

Goal: Integrate a novel object detection module to transform implicit visual information into explicit, structured knowledge, aiming to achieve significant improvement over the established 53% baseline and set a new standard for performance on critical perspective-change scenarios.

3. ORIGINAL MODEL

Model UsedL: llava-onevision-7b
Model Type: Vision-Language Model (VLM)
Task: Multimodal Retrieval-Augmented Generation (MRAG) evaluation, multiple-choice (A/B/C/D) format
Baseline Accuracy: 58.2% (Achieved on MRAG-Bench)



4. MODEL REPLICATION

The system was optimized for hardware constraints while maintaining high performance:

- **Quantization: None**, but **4-bit quantization** could be used to minimize GPU memory usage.
- **Hardware Target:** Optimized to run within a **16GB VRAM** constraint.
- **Memory Footprint:** lava-onevision-7b requires approximately **~16.0 GB** of GPU memory.
- **Inference Settings:** Generation is configured for deterministic output using a **low temperature (0.1)** and greedy decoding (do_sample: false).
- **Input Data:** The model uses **3 of 5** available Ground-Truth (GT) images per question from the HuggingFace MRAG-Bench dataset.

5. ENHANCED PIPELINE

The enhanced pipeline integrates Object Detection to transform image content into structured text, providing **explicit visual grounding** to improve the LLaVA model's reasoning capabilities.

a. Object Detection (DETR):

- The ObjectDetector module uses the **DETR (facebook/detr-resnet-50)** model to analyze each image.
- It extracts labels, confidence scores, and bounding boxes, providing a basis for spatial reasoning.

b. Structured Text Generation:

- Detections are converted into a concise, natural language analysis.
- *Example Output:* "Image 1: Main objects: dog, grass. Detected: 1 dog, 2 grass. Layout: dog in center."

c. Prompt Enhancement:

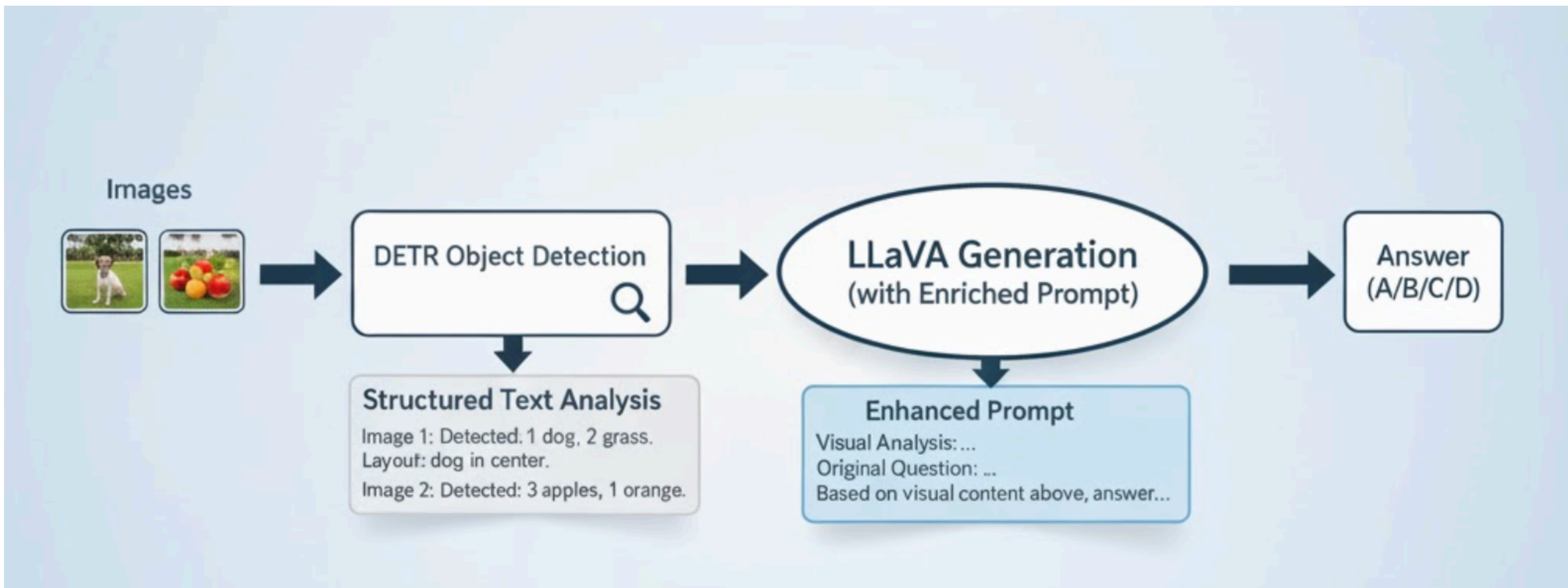
- The structured visual analysis is prepended to the original multiple-choice question, creating an **Enhanced Prompt** that forces **explicit visual grounding**.

d. LLaVA Generation:

- The EnhancedLLaVAPipeline processes the raw images and the enriched prompt to generate the final answer (A/B/C/D) with improved reasoning.

e. Bounding-Box Visual Overlay (Prompt-agnostic):

- What: Overlay DETR boxes/labels on images; prompt unchanged.
- Why: Adds spatial grounding via visuals without extra tokens.
- How: Color-coded rectangles + labels; full metadata (label/confidence/bbox) logged.
- Impact: Compared runs with/without overlays; see Results for accuracy/latency.
- Trade-offs: Small preprocessing; possible occlusion/visual shift. Use alone or with text grounding based on needs.



6. RESULTS

On MRAG-Bench (n=1353), the baseline (Object Detection: DISABLED) achieved 58.2% accuracy (788/1353), meeting the 53–59% target range. Incorporating structured detection text into the prompt produced identical results to the baseline (no measurable change). Using bounding-box overlays on the images (Object Detection: ENABLED) yielded 56.3% accuracy (762/1353), a −1.9 percentage-point change relative to baseline, while still within the target range. Scenario-wise, the overlay condition improved Deformation to 62.7% versus 59.8% (+2.9 pp), remained unchanged on Incomplete (19.6%), Others (60.0%), and Temporal (58.4%), and decreased accuracy on Angle (59.0% vs 64.6%, −5.6 pp) and Partial (60.6% vs 63.8%, −3.2 pp), with smaller declines on Biological (52.9% vs 53.9%, −1.0 pp), Obstruction (63.0% vs 63.9%, −0.9 pp), and Scope (56.9% vs 57.8%, −0.9 pp). Detection produced 6,880 total objects (≈5.1 per sample) with low overhead (~0.12s/sample on average), and resource usage remained within a 16GB VRAM budget. Overall, while object detection consistently provided rich spatial metadata, these inference-time integrations (text augmentation and visual overlays) did not deliver aggregate accuracy gains over the baseline in this setting.

7. CONCLUSION

The Object Detection Enhancement successfully integrates the DETR model to provide structured visual analysis; however, in our experiments, enriching the LLaVA prompt with detection text did not change accuracy, and adding bounding-box overlays decreased accuracy by 1.9 percentage points relative to the 58.2% baseline. While this straightforward inference-time integration did not yield gains over the baseline, the detection outputs remain valuable for diagnostics and future grounding strategies that more directly align with the model's input assumptions.

Future Work:

- **Semantic Segmentation:** Integrate pixel-level segmentation for finer-grained context.
- **Relationship Detection:** Analyze and describe interactions between detected objects to enhance relational reasoning.
- **Multi-scale Detection:** Explore improvements for better detection of small or distant objects.
- **Model Scaling:** Test larger models like LLaVA-1.5-13B or LLaVA-OneVision if additional VRAM is available.