Week 2 Presentation

Charlie Levy

June 2nd 2026

University of Central Florida

Background

ABOUT ME

Year: **CS Rising Junior**

School: **Boston University**

Research: **First project - this one!**

COURSES COMPLETED

- Analysis of Algorithms
- Geometric Algorithms
- Combinatoric Structures
- Probability in Computing
- Computer Systems
- Database Systems

Chosen Project:

Evaluating the Robustness of MLLM Safety Alignment under Image Corruptions

Adeel Yousaf

James Beetham



What's this project about?

The Problem

Current safety evaluations assume clean visual inputs. Real-world images often have blur, noise, etc.

The Approach

Extend existing MLLMs unsafe datasets, add noise, then evaluate safety.

The Goal

How robust are safety-aligned MLLMs under corrupted visual inputs?

Weekly Plan

Week 1: Review literature on MLLM safety and image corruption robustness.

Weeks 2-3: Setup corruption evaluation pipeline including models, datasets, and evaluators.

Weeks 4-7: Evaluate safety under different corruption settings, analyze unsafe compliance and over-refusal behavior, and study robustness patterns across models, corruption types, and training/test-time safety methods.

Week 8: Finalize experiments and prepare paper/report for WACV-2027 submission.

What I've Started

Reading MLLM Literature

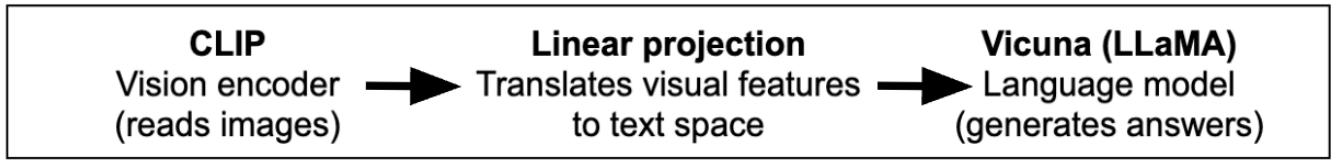
Running LLaVA Locally

MLLMs

[1] Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

Visual Instruction Tuning [1]

- The introduction of LLaVA
- Feed text descriptions of images to GPT-4
- Ask to create realistic Q&A conversations about the images
- Use this data to fine-tune a VLM
- An image gets fed through CLIP, projected, and Vicuna generates a response



Running LLaVA

```
charlielevy@Charlies-MacBook-Air ~/Desktop/REU/llava
❯ python3 llava_chat.py
Running LLaVA locally ('exit' to exit)
To include an image: 'image:/path/to/image.jpg'
System prompt: You are a helpful assistant.

You: Hi

LLaVA:  Hello! How can I assist you today?

You: What's in this image? image:/Users/charlielevy/Downloads/avocado.jpg

LLaVA:  The image shows an avocado. It looks like it could be ripe, as ind

You: exit
charlielevy@Charlies-MacBook-Air ~/Desktop/REU/llava
❯
```



Thank you

Questions?