REPORT - ZERO SHOT

Raghavendra Dabral | +918297139711 | raghavendradabral@gmail.com

This code implements a real-time zero-shot object detection system using OpenAI's CLIP model that identifies objects through semantic similarity rather than traditional bounding boxes. The system processes live webcam feed or video files, analysing each frame to detect objects matching customizable text categories while automatically filtering out common COCO dataset classes. It displays detection results as on-screen text overlays showing the top matching categories with confidence scores. The implementation includes interactive keyboard controls for dynamically updating detection categories during runtime, JSON-based logging that saves results with timestamps to detections_log.json, and GPU acceleration when available. Using a threaded architecture, it maintains responsive performance even during model updates, with configurable parameters for detection threshold and processing speed optimized for real-time operation.

There were many challenges during development. Initially, using Grounding DINO was considered for its performance and support for both bounding boxes and confidence scores. However, setup for Linux Nvidia drivers was taking too long to configure, so sticking to CLIP with better frame rate was the target. Trying to use Gradio significantly affected the FPS, hence, the decision was made to stick to using terminal. Then, the live prompting feature was crashing the program many times, and later, the prompted categories weren’t getting logged into the list of classes. Eventuall,y these concerns were fixed, and the code delivers all of the requirements with some bonus requirements as well. Future improvements include having a more dynamic UI which keeps the frame rate up, and using other models that allow for compatible configuration with Linux Nvidia Cuda drivers 12.1+.