# **Zero-Shot Object Detection - Write-up**

## **How it works**

This project implements a real-time zero-shot object detection system using OWL-ViT, a vision-language transformer model. The user provides custom prompts (labels) via a simple interface, and the model detects those objects in a live webcam feed. Detections are visualized with bounding boxes and labeled with confidence scores, while also being logged into a CSV file with timestamps. The GUI allows users to add or remove prompts dynamically, and visual feedback is provided directly on the video stream.

## **Challenges faced**

Key challenges included optimizing real-time performance while using a large transformer-based model, ensuring smooth frame rates during inference, and balancing detection frequency with responsiveness. Another difficulty was handling false positives — for instance, detecting sunglasses even when not worn — which points to the sensitivity and generalization limits of zero-shot models.

## **What could be improved or added next**

Future improvements could include multi-threaded processing for better performance, adding region-of-interest selection to focus on specific areas, and allowing video file input. Enhancements to the prompt system like natural language queries or integration with speech-to-text could make the system even more interactive. Additionally, model fine-tuning or ensemble methods might help reduce false detections.