Introduction:

Object detection is a pivotal task in computer vision for identifying and localizing objects on images, videos seamlessly and has seen remarkable advancements over the past decade. The advancements of deep learning revolutionized the field, with Convolutional Neural Networks (CNNs) advancements architectures such as R-CNN, YOLO (You Only Look Once), and SSD (Single Shot MultiBox Detector) played significant roles in enhancing detection performance, speed, and scalability.

Object detection models effectively bridged the gap between classification and segmentation by simultaneously identifying the category and precise location of objects on images. Unlike classification models, which only determine what is in an image and segmentation models which focus on pixel-level accuracy, computationally intensive. On the other hand, object detection provided a more efficient and scalable solution by generating bounding boxes around objects and identifying their categories. This dual capability allows for real-time processing and higher frames per second (FPS), making it suitable for applications that require immediate responses, such as autonomous driving, surveillance, and real-time video analysis. Additionally, it significantly reduces the cost burden compared to segmentation, enabling it to be used in various edge devices and less hardware-specific applications.

Object detection, while powerful, faces several limitations that have opened the door for vision-language models to enhance and broaden its capabilities. Traditional object detection models struggle with accuracy in complex scenes containing many overlapping objects or cluttered backgrounds, leading to misclassifications and localization errors. Additionally, detecting small objects within images remains challenging due to limited distinctive features, often resulting in lower accuracy compared to larger objects.