# Zero-Shot Aerial Traffic Accident Detection: A Comparative Study of OmDet-Turbo and Moondream2

Rosmarino Fabrizio
Department of Computer Science
University of Bari
f.rosmarino@studenti.uniba.it

June 4, 2025

**Abstract**

This study explores the application of open-vocabulary object detection techniques for the automatic identification of traffic accidents from aerial imagery. The objective is to assess the capability of vision-language models to operate in unsupervised scenarios, where object categories are not predefined but specified at runtime through textual descriptions (prompts). Two open-source models were selected: OmDet-Turbo, a real-time detector based on the DETR architecture, and Moondream2, a model oriented towards multimodal understanding. To support the evaluation, a custom dataset comprising 1000 images was constructed, half of which were obtained from public video frames and the other half generated artificially. Each image was manually annotated with bounding boxes in normalized YOLO format, indicating the presence of crashed vehicles. The models were tested in zero-shot mode, using semantic prompts such as "identify severely damaged cars." Predictions were standardized in YOLO format to facilitate evaluation. The assessment was conducted both qualitatively, through visualization of overlaid boxes, and quantitatively using Intersection over Union (IoU) and mean Average Precision (mAP) metrics. Results indicate that OmDet-Turbo achieves superior spatial accuracy, making it suitable for real-time scenarios. Moondream2, while less precise, proves useful in low-resource environments due to its simplicity and ability to operate without advanced hardware. Experimental notebooks are available online and can be accessed through the project's GitHub repository.

## 1 Introduction

Timely identification of traffic accidents is crucial for enhancing urban safety, emergency response, and infrastructure monitoring. In this context, intelligent drones offer a unique perspective by providing real-time aerial imagery, potentially identifying critical events before human intervention. This work addresses the problem of automatic detection of traffic accidents in aerial images through open-vocabulary object detection techniques. Unlike conventional approaches based on a fixed set of known classes (closed-set), the aim is to test models capable of recognizing objects described textually, even if never seen during training—a mode known as zero-shot detection. To explore this perspective, two open-source models with different architectures and purposes were selected: OmDet-Turbo, optimized for real-time applications and equipped with a DETR architecture enhanced by an Efficient Fusion Head, and Moondream2, a vision-language model suitable for devices with limited computational power. Both models were tested on a custom dataset of 1000 images, manually annotated with bounding boxes related to crashed vehicles. The research objective is twofold: on one hand, to evaluate the effectiveness of these models in zero-shot detection in realistic scenarios; on the other, to provide a systematic comparison in terms of spatial and semantic accuracy, using metrics such as IoU and mean Average Precision (mAP), thus offering concrete insights for the application of AI technologies in automated urban monitoring.

## 2 Related Work

### Closed-Set and Open-Vocabulary Object Detection

Traditional object detection relies on closed-set models, where the object classes to be detected are fixed during training. YOLO, SSD, and Faster R-CNN are emblematic examples of this family. However, these approaches are limited in dynamic real-world scenarios, where new or rare categories may appear. In recent years, open-vocabulary object detection (OVD) models have emerged, allowing the detection of categories unseen during training, guided by textual descriptions. A major contribution in this field is GLIP (Grounded Language-Image Pretraining) by Li et al. [1], which integrates vision-language learning for prompt-based detection. Similarly, OWL-ViT extends this approach to Vision Transformers, enabling multimodal matching and detection in zero-shot mode [2].

### Open-Vocabulary on Aerial Images

The application of OVD techniques to aerial images is relatively recent but growing rapidly. Wei et al. proposed OVA-DETR [3], which combines bidirectional vision-language fusion with a DETR-based architecture to detect objects in complex UAV-based scenarios. The model aligns visual semantics with textual descriptions, enhancing generalization to rare categories. In a different approach, Li et al. introduced CastDet [4], which leverages CLIP as a teacher model to generate pseudo-labels in open-vocabulary contexts. This is particularly useful when explicit annotations are scarce and there is a need to expand the detectable vocabulary.

### Accident Detection from Drones

In the more specific domain of road accident detection from aerial images, Boddu and Mukherjee [5] applied YOLOv5 to identify ambulances and crashed vehicles. Although this is a closed-set model, it showed promising results in terms of speed and efficiency in emergency scenarios. More recently, Li et al. (2023) [6] combined open-vocabulary techniques with semantic segmentation to enhance object recognition in UAV-based disaster response contexts. Their method integrates linguistic embeddings to dynamically adapt to categories not present in the training data.

### Positioning of This Study

This work situates itself within this research direction by proposing an experimental comparison between two open-source vision-language models applied to the detection of traffic accidents in aerial images: OmDet-Turbo, a true open-vocabulary object detector, and Moondream2, a multimodal model designed for general visual understanding tasks. Although the two models differ in architecture and intended use, the comparison aims to assess their ability to operate in zero-shot scenarios, evaluating both the spatial coherence of their predictions and their semantic interpretability in realistic urban contexts.

## 3 Dataset

To enable a comparative evaluation of the OmDet-Turbo and Moondream2 models, it was constructed a dedicated dataset specifically for this study, focused on the detection of road accidents and the identification of damaged or crashed vehicles. The dataset comprises approximately 1000 aerial images, developed from scratch through a mix of manual collection and artificial generation. Out of the total, I personally curated and annotated around 200 images, while the remaining samples were produced in collaboration with other contributors. The images were obtained through two main methods:

- Frames extracted from publicly available videos (e.g., YouTube) or images sourced from the web, depicting real-world urban traffic accidents;
- Artificially generated scenes created using graphical tools, intended to enhance variability and simulate less common conditions (e.g., aerial perspectives, poor lighting, presence of debris or emergency responders).

To align with the intended drone-based application scenario, all images were standardized to a resolution of 1280 × 720 pixels, matching the typical output of onboard drone cameras. Images were resized prior to annotation to ensure consistency across the dataset.

Each image was paired with an annotation file in YOLO format, where each detected object is represented using the following syntax:

```
<class_id> <x_center> <y_center> <width> <height>
```

All coordinates are normalized relative to the image dimensions. For this study, I used a single class (`class_id = 0`) to indicate the presence of a *crashed vehicle*, allowing the focus to remain on the models' semantic understanding capabilities in a zero-shot setting.

All bounding boxes were manually annotated using the LabelImg tool to ensure accuracy and semantic alignment with the scene. I chose the YOLO format due to its broad compatibility with modern object detection frameworks and its efficiency in the evaluation pipeline.

**Note.** The dataset was used exclusively for inference and evaluation. Neither OmDet-Turbo nor Moondream2 was fine-tuned on this dataset, ensuring that all results reflect a genuine zero-shot detection scenario.

# 4 Methods

## OmDet-Turbo

OmDet-Turbo is an open-vocabulary object detection model designed for real-time applications. Derived from the DETR (DEtection TRansformer) framework, it adopts a Transformer-based encoder-decoder architecture enhanced with an efficient language-vision fusion module known as the Efficient Fusion Head (EFH).

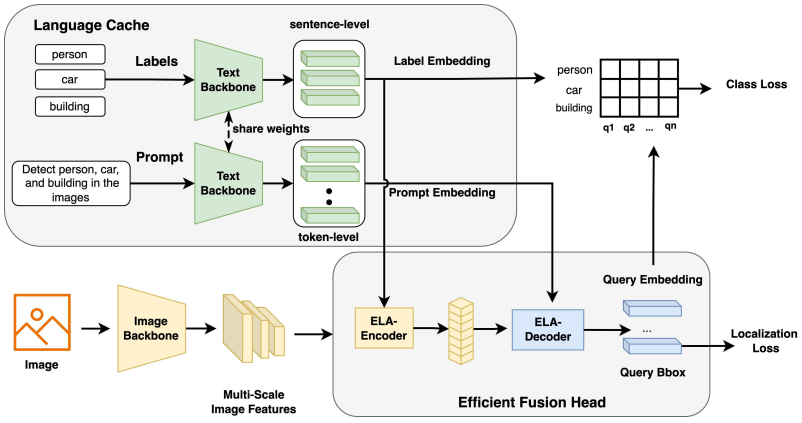


Figure 1: OmDet-Turbo model architecture

The architecture consists of three main components. The textual module is a Transformer-based language encoder that processes both prompts (e.g., "detect crashed cars") and class labels. Prompts are encoded at the token level to retain fine-grained semantics, while labels are embedded at the sentence level. The visual backbone, based on a convolutional or Vision Transformer architecture such as ConvNeXt, extracts a multiscale feature pyramid (P3, P4, P5) from the input image. The EFH module, composed of an encoder and a decoder, selects semantically relevant queries and combines them with the language prompt to generate final predictions (classes and bounding boxes). Key features include pre-caching of text embeddings, scalability to large vocabularies, and flexibility in handling dynamic prompts. These capabilities enable OmDet-Turbo to operate in zero-shot mode, detecting objects described in the prompt even if they were not seen during training. The model was pretrained on four types of datasets: generic object detection (e.g., "detect objects" with labels like "car"), grounding (e.g., "a laptop is on the table" linked to a visual label), visual question answering (questions with visual answers), and human-object interaction (e.g., "a person holding a bottle") [7].

## Moondream2

Moondream2 is an open-source vision-language model optimized for deployment on resource-constrained devices. It includes approximately 1.86 billion parameters and consists of two core modules: SigLIP as the visual encoder and Phi-1.5 as the language model.

SigLIP replaces the traditional softmax loss (used in CLIP) with a pairwise sigmoid loss, which improves performance in zero-shot classification and image-text retrieval tasks. Operating exclusively on image-text pairs, it allows for smaller and more efficient batch processing [8].Phi-1.5, developed by Microsoft, is a compact Transformer-based language model with 1.3 billion parameters. Trained on 30 billion tokens—including 7 billion from Phi-1 and around 20 billion synthetically generated by GPT-3.5—it has demonstrated competitive results across a variety of language understanding tasks, outperforming much larger models [9].Moondream2 performs well in lightweight multimodal tasks such as visual question answering (VQA), image captioning, OCR, and spatial coordinate extraction. While not originally designed for object detection, it includes a `detect()` interface that allows for approximate object localization based on textual descriptions. In zero-shot mode, it can generate estimated bounding boxes for unseen objects without requiring additional training. However, the accuracy is lower compared to specialized detectors and strongly depends on prompt formulation [11]. Despite these limitations, Moondream2 is a valuable tool in scenarios where model efficiency and semantic flexibility outweigh spatial precision.

## Architectural and Functional Comparison

After detailing the architectures of OmDet-Turbo and Moondream2, it is useful to compare their operational characteristics in relation to the task at hand. Although developed with different objectives, both models were applied to the zero-shot object detection setting, revealing unique strengths and trade-offs.OmDet-Turbo is a specialized system designed for prompt-guided detection, with a pipeline tailored to efficiently integrate visual and linguistic information. Moondream2, although not explicitly developed for this task, demonstrates baseline detection capabilities thanks to its multimodal structure and dedicated inference interface. This flexibility makes it a compelling solution in scenarios where simplicity and portability are key.The following table summarizes the key differences between the two models:

| **Aspect** | **OmDet-Turbo** | **Moondream2** |
|---|---|---|
| Parameters | ∼100M | ∼2B |
| FLOPs | ∼30 GFLOPs | ∼100 GFLOPs (estimated) |
| Architecture | Transformer + Efficient Fusion Head | SigLIP (visual) + Phi-1.5 (language) |
| Primary Purpose | Open-vocabulary object detection | General vision-language understanding |
| Bounding Box Quality | Precise and semantically consistent | Approximate, sensitive to input formulation |
| Zero-shot Support | Extensive, with fine-grained semantic control | Partial, less specialized |
| Hardware Compatibility | Optimized for GPU (real-time) | Efficient on CPU and edge devices |
| Strengths | Accuracy, speed, semantic prompt flexibility | Versatility, lightweight, ease of use |
| Limitations | Higher integration complexity | Not designed for structured detection |

Table 1: Comparison between OmDet-Turbo and Moondream2 in terms of architecture and usability. Data derived from [7] for OmDet-Turbo, and from [9, 11] for Moondream2.

In summary, the analysis highlights that OmDet-Turbo achieves superior performance in terms of spatial accuracy and semantic control, making it particularly suitable for complex, real-time scenarios where detection precision is critical. Moondream2, on the other hand, while not specifically built for object detection, shows good adaptability to the task in less demanding settings—thanks to its lightweight architecture and ability to function on low-resource hardware.

# 5 Experiments

This section presents the experimental tests I conducted on the OmDet-Turbo and Moondream2 models, both applied in a *zero-shot object detection* context—that is, without any fine-tuning, on the custom dataset built specifically for this study. The goal was to detect crashed or damaged vehicles in aerial images, simulating a realistic scenario of automated urban monitoring using drone imagery.

Although sharing the same final objective, the two models rely on different approaches:

- **OmDet-Turbo** natively supports open-vocabulary detection guided by textual prompts. I tested it using semantically rich input descriptions combined with the relevant target classes.
- **Moondream2**, by contrast, operates solely based on class labels due to its architecture. It does not support detection based on multimodal prompts and cannot return bounding boxes directly guided by natural language. I therefore tested it exclusively by providing isolated semantic class names, without contextualized descriptions.

## Prompts and Test Classes

Three descriptive prompts were used to guide inference in **OmDet-Turbo**:

- `"Detect all Crashed cars which are completely crushed."`
- `"Detect all damaged cars."`
- `"Detect all Car accident which refers to the area of a crash, including debris or emergency vehicles."`

These prompts correspond to three distinct semantic categories: *Crashed car*, *Damaged car*, and *Car accident*. I used the same three classes for **Moondream2**, which, however, was limited to simple label-based detection.

## Prediction Format

To standardize the evaluation process, I converted all predicted bounding boxes to normalized YOLO format. Since the native output formats differ between the two models, specific transformations were required.

**OmDet-Turbo** produces bounding boxes in absolute coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$, which I converted using the following formulas:

$$x_{\text{center}} = \frac{x_{\min} + x_{\max}}{2W}, \quad y_{\text{center}} = \frac{y_{\min} + y_{\max}}{2H}$$
$$\text{width} = \frac{x_{\max} - x_{\min}}{W}, \quad \text{height} = \frac{y_{\max} - y_{\min}}{H}$$

**Moondream2**, on the other hand, returns normalized bounding boxes in the format `x_min`, `y_min`, `x_max`, `y_max`. I converted these into YOLO-centered format using:

$$x_{\text{center}} = \frac{x_{\min} + x_{\max}}{2}, \quad y_{\text{center}} = \frac{y_{\min} + y_{\max}}{2}$$
$$\text{width} = x_{\max} - x_{\min}, \quad \text{height} = y_{\max} - y_{\min}$$

## Qualitative Evaluation

Before conducting the quantitative evaluation, I carried out an initial qualitative analysis to explore how each model behaves when applied to real aerial imagery. This phase provided early insights into detection accuracy, semantic alignment, and typical failure modes.

The process involved automatically overlaying the predicted bounding boxes on the original images for visual inspection. I implemented a script to generate these visualizations in batch mode, processing all images in the dataset using the model outputs. This automation ensured consistency, efficiency, and reproducibility throughout the assessment.

Each prediction was evaluated according to the following criteria:

- **Spatial accuracy**: how precisely the bounding boxes frame the actual vehicles;
- **Semantic consistency**: particularly for OmDet-Turbo, whether the detections reflect the intended meaning of the prompts;
- **Error patterns**: including false positives (e.g., detecting intact or unrelated vehicles) and false negatives (missed detections of actual crashes).

Figures 2 and 3 display representative outputs from each model, selected to illustrate the general characteristics of their detection performance.


Figure 2: OmDet-Turbo output with the prompt: `"Detect all damaged cars"`.

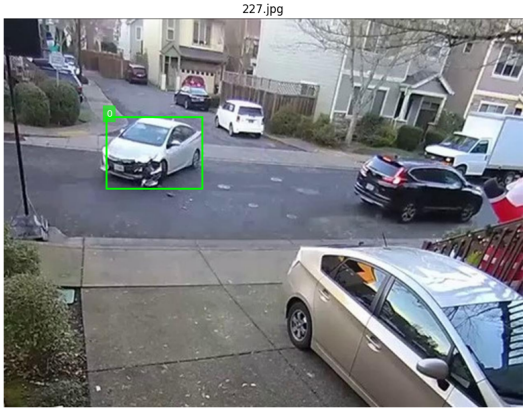


Figure 3: Moondream2 output using the class label `"damaged car"`.

Across the dataset, OmDet-Turbo generally produced more precise and semantically coherent detections than Moondream2. Its bounding boxes were often well-aligned with vehicles involved in crashes, and the model showed a solid ability to interpret linguistic prompts by focusing on relevant visual cues such as deformations, displacements, or scene context. Nonetheless, some detections were imprecise or missed entirely, particularly under low visibility or in visually complex scenes.

Moondream2, by contrast, generated fewer detections overall. However, the bounding boxes it produced were often correct with respect to the target class, even if less precise in their placement compared to those of OmDet-Turbo. The model performed well in simple scenes but struggled in cluttered or ambiguous environments, suggesting a more limited understanding of the visual context.

## Quantitative Evaluation

To complement the qualitative analysis, I also conducted a quantitative evaluation using two widely adopted object detection metrics:

- **Intersection over Union (IoU)** — measuring the overlap between predicted and ground-truth bounding boxes;
- **mean Average Precision (mAP)** — computed at IoU thresholds from 0.50 to 0.75 to assess both detection accuracy and robustness.

All evaluations were performed using the normalized YOLO-format annotations, ensuring consistency between ground truth and model outputs.

**OmDet-Turbo Results (prompt + classes)**

| Class | IoU | mAP |
|---|---|---|
| Crashed car | 0.66 | 0.48 |
| Damaged car | 0.77 | 0.21 |
| Car accident | 0.74 | 0.36 |

Table 2: OmDet-Turbo performance on semantic prompts.

**Moondream2 Results (class labels only)**

| Class | IoU | mAP |
|---|---|---|
| Crashed car | 0.42 | 0.57 |
| Damaged car | 0.47 | 0.51 |
| Car accident | 0.43 | 0.37 |

Table 3: Moondream2 performance on class labels.

The quantitative analysis shows that OmDet-Turbo and Moondream2 exhibit different strengths in the context of traffic accident detection from aerial images. The choice of the most suitable model depends closely on the operational requirements of the UAV system, particularly the balance between spatial accuracy and the computational capacity available on board.

**OmDet-Turbo** achieved the highest *Intersection over Union (IoU)* values across all analyzed classes, confirming a remarkable ability in spatial localization. This feature is essential in real scenarios, where a drone must accurately identify the areas involved in an accident, damaged vehicles, debris, or other relevant contextual elements. Moreover, the ability to use articulated textual prompts gives the model greater semantic flexibility, useful in adaptive or dynamic missions.

**Moondream2**, while showing lower performance in terms of localization (IoU), achieved higher *mean Average Precision (mAP)* values in the *Crashed car* and *Damaged car* classes. This indicates a greater ability to detect objects with sufficient confidence, albeit with less precise positioning. Its lightweight structure and CPU-compatible execution make it suitable for low-power UAV platforms, where GPU usage is not feasible.

# 6 Conclusion

This study compared OmDet-Turbo and Moondream2 for zero-shot traffic accident detection in aerial imagery, with the goal of supporting autonomous UAV-based monitoring without task-specific training. OmDet-Turbo showed superior spatial precision and a strong ability to interpret complex prompts, making it ideal for high-accuracy real-time scenarios.Moondream2, while generating fewer detections and showing lower localization performance (as reflected by lower IoU scores), achieved higher mean Average Precision (mAP) values in the *Crashed car* and *Damaged car* categories. This outcome may seem counterintuitive, as mAP is often associated with accurate localization. However, mAP also rewards predictions that correctly classify objects with high confidence, even if their bounding boxes are not perfectly aligned. In practice, this means that Moondream2 can achieve good mAP scores by reliably recognizing the correct class—especially in simple scenes—even when the bounding boxes are slightly off. These results underline a practical trade-off between detection accuracy and deployment flexibility, suggesting that the two models can serve complementary roles depending on operational constraints. Future work could extend this framework in several directions. First, incorporating temporal information would allow the system to reason over video streams and track accident dynamics over time. Second, hybrid detection pipelines could be explored, with Moondream2 acting as a lightweight region proposal module and OmDet-Turbo refining results via semantic prompts. Since OmDet-Turbo does not support fine-tuning, enhancing detection performance may depend on prompt engineering strategies that dynamically adapt to scene context. Additional robustness could be achieved by fusing data from multiple modalities, such as GPS, IMU, or thermal sensors, especially under low-visibility

conditions. Lastly, compression techniques like quantization or pruning could make Moondream2 even more suitable for real-time inference on embedded UAV hardware.

# References


[1] Li, Z. et al. "Grounded Language-Image Pretraining." CVPR 2022. https://arxiv.org/abs/2112.03857

[2] Minderer, N. et al. "Simple Open-Vocabulary Object Detection with Vision Transformers." CVPR 2022. https://arxiv.org/abs/2205.06230

[3] Wei, Z. et al. "OVA-DETR: Open-Vocabulary Aerial Object Detection with Bidirectional Language-Vision Alignment." arXiv 2024. https://arxiv.org/abs/2408.12246

[4] Li, Z. et al. "CastDet: Category-Aware Student-Teacher Detection for Open-Vocabulary Object Detection." arXiv 2023. https://arxiv.org/abs/2311.11646

[5] Boddu, S. and Mukherjee, S. "Detection of Emergency Vehicles and Road Accidents in Aerial Imagery using YOLOv5." arXiv 2024. https://arxiv.org/abs/2412.05394

[6] Li, Z. et al. "UAV-Based Open-Vocabulary Object Detection for Disaster Response." Drones 2023. https://www.mdpi.com/2504-446X/9/2/155

[7] Jiancheng Zhao et al., *Real-time Transformers-based Open-Vocabulary Detection with Efficient Fusion Head*, arXiv:2403.06892. https://arxiv.org/abs/2403.06892

[8] Touvron, H., et al. (2023). *SigLIP: Scaling up Vision-Language Pretraining with Pairwise Sigmoid Loss*. arXiv preprint arXiv:2306.00989. https://arxiv.org/abs/2306.00989

[9] Gunasekar, S., et al. (2023). *Textbooks Are All You Need II: phi-1.5 technical report*. arXiv preprint arXiv:2309.05463. https://arxiv.org/abs/2309.05463

[10] Data Science Dojo. (2024). *Vision Language Models: Introducing the new VLM Moondream 2*. https://datasciencedojo.com/blog/vision-language-models-moondream-2/

[11] Vikhyat et al., *moondream GitHub repository (2024)*. https://github.com/vikhyat/moondream

[12] Carion et al., *End-to-End Object Detection with Transformers*, FAIR, 2020. https://arxiv.org/abs/2005.12872