Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
6(2): 115–137.
DOI: 10.3934/aci.2026007
Received: 07 May 2026
Revised: 05 July 2026
Accepted: 10 July 2026
Published: 20 July 2026

Research article

Cross-paradigm evaluation of gaze-based object identification for human-centered vehicle interfaces

Penghao Deng, Jidong J. Yang* and Jiachen Bian

Smart Mobility & Infrastructure Laboratory, College of Engineering, University of Georgia, Athens GA 30602, USA

* **Correspondence:** Email: Jidong.Yang@uga.edu.

Academic Editor: Zhiling Long

Abstract: Understanding where drivers direct their visual attention during driving, known as gaze behavior, is important for developing next-generation advanced driver-assistance systems and improving road safety. This paper addressed this challenge as a semantic identification task from road scenes captured by a vehicle's front-view camera. Specifically, the collocation of gaze points with object semantics was investigated using three distinct vision-based approaches: direct object detection, segmentation-assisted classification, and query-based vision-language models. The results demonstrate that the direct object detection and vision language model–based approach outperformed other approaches, achieving macro F1-scores exceeding 0.84. The larger vision-language model, in particular, exhibited better robustness and performance for identifying small, safety-critical objects such as traffic lights, especially in nighttime conditions. Conversely, the segmentation-assisted paradigm suffers from a *part-versus-whole* semantic gap, which led to low recall. The results reveal a trade-off between the real-time efficiency of traditional detectors and the contextual understanding and robustness offered by large vision-language models. These findings provide guidance for the design of future human-aware driver monitoring systems.

Keywords: gaze object identification; advanced driver-assistance systems (ADAS); vision-language models (VLMs); object detection; driver attention

1. Introduction

Despite decades of technological advancements in vehicle engineering and infrastructure, ensuring road safety remains a major global challenge. Each year, road traffic incidents result in a large number of fatalities and injuries. The U.S. National Highway Traffic Safety Administration projected 39,345 traffic fatalities in the United States for 2024, a figure that, while representing a slight decrease,

remains high compared to levels from a decade prior [1]. On a global scale, road traffic injuries are the leading cause of death for individuals aged 5 to 29, with an estimated 1.19 million lives lost annually [2]. A significant contributor to this problem is driver inattention, a multifaceted issue encompassing any activity that diverts a driver's focus from the primary task of safely operating a vehicle [3,4]. Driver distraction, a subset of inattention, is implicated in a substantial portion of police-reported crashes, with estimates suggesting its involvement in 8% of fatal crashes and 13% of injury crashes in 2023 alone [5].

Naturalistic driving studies have provided quantitative evidence linking inattention to increased crash risk. Research has shown that engaging in visually or manually complex secondary tasks can increase the near-crash or crash risk by a factor of three compared to attentive driving [6]. In addition, a single glance away from the forward roadway lasting more than two seconds has been found to increase this risk by at least twofold [7]. The proliferation of in-vehicle infotainment systems and personal mobile devices has increased this problem, introducing additional sources of visual, manual, and cognitive distraction into the cockpit [8].

In response to this safety issue, the automotive industry and regulatory bodies have promoted the development and deployment of driver monitoring systems (DMS). These systems utilize in-cabin sensors and artificial intelligence to assess a driver's state in real time, detecting conditions such as drowsiness, distraction, and impairment [9]. More broadly, recent studies have shown that machine learning–based methods can effectively leverage monitored transportation data to support detection, prediction, and decision-making under complex real-world conditions[10,11]. While current DMS technologies are effective at identifying overt signs of inattention, such as prolonged eye closure or cell phone use, a remaining challenge is achieving a more detailed understanding of the driver's situational awareness. This requires moving beyond simple distraction detection to proactively assessing the driver's perception of the dynamic environment.

Central to this task is the analysis of the driver's gaze. A driver's point of gaze is a direct indicator of their visual attention, providing information about their cognitive state, intentions, and awareness of surrounding traffic elements [12,13]. By precisely identifying the semantic object at which a driver is looking, it becomes possible to build human-centric vehicle systems. The applications of this capability can be extensive and transformative.

First, in the domain of advanced driver-assistance systems (ADAS), it enables the creation of context-aware warning systems. An ADAS that knows a driver has failed to visually register an approaching pedestrian can issue a timely alert, whereas a system that confirms the driver is already monitoring the hazard can suppress a redundant and potentially annoying warning, thereby reducing alarm fatigue [13]. This shifts the focus from vehicle-centric perception, where the system only knows what it sees, to human-centric perception, where the system understands what the driver sees.

Second, for the development of human-centered autonomous vehicles, analyzing human gaze patterns provides useful data for designing more intuitive and trustworthy artificial intelligence (AI) driving policies [14]. During safety-critical handover events in semi-autonomous vehicles, confirming that the driver's attention is appropriately directed toward the relevant hazard is important for ensuring a safe transition of control [15,16]. Furthermore, gaze behavior serves as a robust, implicit measure of a driver's trust in the automated system, a key factor in technology acceptance and safe operation [17].

Third, this technology can improve in-vehicle human–machine interfaces (HMIs). HMIs can be designed to adaptively present information in the driver's line of sight or to use subtle cues to guide attention toward critical information, improving situational awareness without increasing cognitive load [18].

Finally, for driver behavior analysis, such as in post-accident forensics, driver training, or insurance telematics, identifying gazed objects provides detailed information about a driver's risk perception and decision-making processes that is not captured by current metrics like speed and acceleration data [19].

Formally, the research problem addressed in this paper is point-of-gaze object identification: given an image frame from a vehicle's forward-facing camera and a corresponding time-synchronized gaze coordinate of the driver on the image plane, the task is to determine the semantic class of the object at that gaze point; in other words, which object in the driving scene the driver is paying attention to at a particular time (frame). Despite its clear definition, this task has several technical challenges that stem from the dynamic nature of real-world driving.

A primary challenge is the variability of driving scenes. The performance of any vision-based system must be robust across a wide spectrum of environmental conditions. Adverse weather, such as rain, snow, and fog, can severely degrade image quality, introducing noise and obscuring object features [20]. Similarly, lighting conditions vary dramatically, from the bright glare of direct sunlight to the low-light and high-contrast environments of nighttime driving, each posing unique difficulties for perception algorithms [21]. The objects of interest themselves, such as pedestrians, vehicles, and traffic signals, exhibit a wide range of appearances, scales, and orientations. They can be partially occluded by other road users or infrastructure, making their complete and accurate identification a difficult problem [22].

Another challenge lies in the inherent ambiguity of a single gaze point. Geometrically, a coordinate is ambiguous. It may fall on a small, distinct object like a distant traffic light, on a specific component of a larger object like the wheel of a truck, or in the empty space between two adjacent vehicles [23]. This ambiguity also has a semantic dimension: a driver looking at a car's taillight is semantically attending to the car as a whole entity, not just the taillight. Any effective system must resolve this *part-versus-whole* relationship to correctly infer the driver's focus of attention. This technical challenge is related to well-documented phenomena in cognitive psychology, such as inattentional blindness and looked-but-failed-to-see (LBFTS) errors [24]. LBFTS incidents occur when a driver's gaze is directed at a hazard, yet they fail to consciously perceive it, often due to a lack of expectation or cognitive overload [25,26]. This shows that the problem is not merely one of geometric localization but of inferring a latent human cognitive state, where the ground truth is the driver's semantic intent.

Finally, for these applications to be practical, particularly those involving real-time warnings or control transitions (e.g., between human driving and automated driving), the entire identification process must be executed with minimal latency. This imposes a limited computational budget on the system, favoring architectures that are both accurate and efficient [27]. The necessity of balancing high accuracy under diverse conditions with the demands of real-time performance makes point-of-gaze object identification a challenging problem involving computer vision, human factors, and automotive engineering.

To address the task of point-of-gaze object identification, several distinct methodological paradigms can be adapted from the broader fields of computer vision and scene understanding. These can be broadly categorized into object detection-based, segmentation-assisted, and approaches based on vision-language models (VLMs).

1.1. Object detection-based approaches

The most direct method involves using off-the-shelf object detectors such as the you-only-look-once (YOLO) family or region-based convolutional neural networks (R-CNNs) [28,29]. These models have been extensively applied to traffic scene understanding and driver monitoring, showing high efficiency in the case of single-stage detectors like YOLO and high accuracy from two-stage detectors like Faster R-CNN [30]. The operational principle is straightforward: if a driver's gaze coordinate falls within a predicted bounding box, the gazed object is identified by that box's class label. However, this approach has several limitations in the context of gaze analysis. First is the *gaze-in-the-gap* problem: the method fails if the gaze point lands on the background or between detected objects, providing no information even if the driver is looking at an undetected target object [31]. Second, standard detectors often struggle with small, distant, or partially occluded objects, which are frequent in driving scenes [32]. Finally, a bounding box provides only a coarse localization, meaning a gaze point could fall near the boundary or a corner of a bounding box, where the gazed pixel belongs to the background rather than the object of interest, leading to an incorrect prediction.

1.2. Segmentation-assisted classification approaches

To address the issue with bounding boxes and achieve pixel-level localization, a two-stage approach can be employed, first using an image segmentation model to isolate the object at the gaze coordinate, followed by a classification model to identify the resulting image crop. Semantic segmentation networks can provide pixel-level segmentation of driving scenes, delineating drivable areas and common object categories with high accuracy [33]. The recent development of foundation models for segmentation, most notably the segment anything model (SAM), offers a useful tool [34]. SAM can generate high-quality segmentation masks for arbitrary objects based on point or box prompts, showing strong zero-shot generalization capabilities [35]. However, this paradigm also has drawbacks. A primary limitation is its class-agnostic segmentation: it produces a mask but provides no semantic label, necessitating a second classification stage [36]. This two-stage pipeline introduces additional computational overhead, compromising real-time performance. This approach is also susceptible to the *part-versus-whole* problem described above. For instance, if a driver's gaze lands on a car's wheel, SAM may segment only the wheel rather than the whole car.

1.3. VLM-based approaches

A recent paradigm involves the use of large VLMs. These models, which are pre-trained on vast datasets of paired images and text, have demonstrated capabilities in joint visual and linguistic reasoning [37]. For the task of gaze object identification, a VLM can be prompted with the full image, the gaze coordinate, and a natural language query such as, "What is the object at coordinate (x, y)?" This reframes the problem as a form of visual question answering (VQA) [38]. Modern open-source VLMs, such as the Qwen-VL series, have shown strong performance in spatial understanding, fine-grained object recognition, and answering complex queries about traffic scenarios [39]. Their ability to perform visual grounding, linking textual concepts to specific image regions, is relevant [40]. This end-to-end, query-based approach can use image context and directly output a semantic label for the task. However, the application of VLMs to this specific, granular problem has not been systematically evaluated in the literature.

1.4. Research gap and contributions

Although these paradigms have different advantages and disadvantages, there is a lack of systematic, comparative studies that evaluate their performance for the specific task of identifying objects at drivers' points of gaze. No unified benchmark exists to assess these different paradigms on equal footing, particularly under diverse environmental conditions in real-world driving. Furthermore, the potential of modern, large VLMs as a direct, query-based solution against established computer vision pipelines for this granular task remains an open question.

This paper aims to address this research gap by presenting a systematic comparison of different vision-based paradigms for identifying the semantic object at a driver's gaze point. In this study, five distinct methods, derived from the three paradigms previously introduced, were evaluated on a benchmark designed to test performance in realistic driving scenarios. The main contributions of this work are as follows:

1) The design and implementation of a comprehensive framework for systematically evaluating and comparing three distinct paradigms (object-detection-based, segmentation-assisted, and VLM-based) for driver gaze–based object identification.
2) An investigation into the application of large VLMs (specifically, Qwen-VL) for this task, providing a direct performance comparison against state-of-the-art computer vision techniques.
3) The development and introduction of a new, manually annotated benchmark dataset based on BDD100K, featuring varied driving conditions (day/night, clear/rainy) to support evaluation under different conditions.
4) A detailed analysis of the performance trade-offs, strengths, and weaknesses of each method across different environmental scenarios and object categories, offering practical insights for developing future driver monitoring and scene perception systems.

The remainder of this paper is organized as follows: Section 2 details our methodology, covering the benchmark dataset, the compared vision paradigms, and the evaluation metrics. Section 3 presents a systematic analysis of the experimental results. Section 4 discusses the key findings, including the practical implications and the practical trade-off between performance and efficiency. Finally, Section 6 concludes the paper with a summary of our findings and outlines directions for future research.

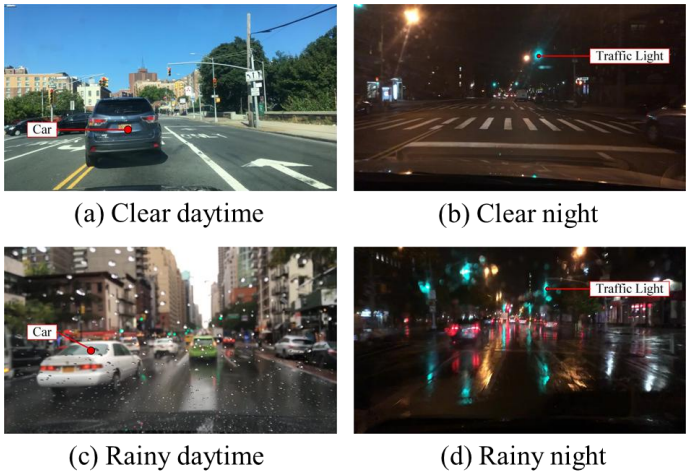
2. Benchmark dataset

To conduct a rigorous evaluation of the different vision-based paradigms, a diverse and realistic benchmark dataset is essential. For this study, a benchmark dataset was created from the BDD100K dataset [41], which covers a wide range of driving scenarios, including a wide variety of weather conditions, times of day, and scene types. These characteristics provide a realistic and challenging basis for assessing the performance and robustness of different methods.

The benchmark dataset was created by selecting images exclusively from the city street scene category. These images were then filtered to construct four distinct and challenging environmental conditions: clear daytime, clear night, rainy daytime, and rainy night. The annotation protocol involved a manual process where points were placed on objects of interest within each selected image. This procedure was designed to simulate driver gaze points. The final dataset is composed of a collection of image-coordinate-label triplets, which serve as the ground truth for all subsequent experiments.

Figure 1 presents representative examples from each of the four annotated scenarios, illustrating the annotation method and the visual complexity of the task. Each scenario presents unique challenges to vision algorithms. The clear daytime condition, shown in Figure 1(a), features good visibility but

can include harsh shadows and glare. The clear night condition in Figure 1(b) is characterized by low ambient light and high contrast from artificial sources such as streetlights and vehicle headlights. In the rainy daytime scenario, shown in Figure 1(c), image quality is degraded by reduced visibility, reflections on wet surfaces, and occlusions from raindrops on the windshield. The rainy night scenario, depicted in Figure 1(d), is most challenging as it combines the difficulties of nighttime driving with the image degradation effects of rain. The red markers shown in each image are examples of the simulated gaze point annotations used for model evaluation.



(a) Clear daytime (b) Clear night

(c) Rainy daytime (d) Rainy night

Figure 1. Representative examples from the four annotated scenarios and annotation protocol.

Five target object classes were selected for this study: person, car, bus, truck, and traffic light. These classes, corresponding to class IDs 0, 2, 5, 7, and 9, respectively, in the original dataset, were chosen based on their prevalence and safety importance in typical urban driving environments. The class IDs align with the COCO dataset [42], providing a standardized basis for evaluation. To ensure a fair and consistent assessment across all methods, a label mapping process was required. This was necessary to harmonize the different class definitions used by the pre-trained models, specifically the COCO-based taxonomy of the YOLOv13 object detector [43] and the ImageNet-1K-based taxonomy of the EfficientNetV2 classifier [44].

Table 1 details the mapping used to align the ImageNet-1K classes with the five broader target categories. The table shows how multiple, more granular ImageNet-1K classes were consolidated into a single target class. For example, ImageNet-1K classes such as beach wagon, cab, and convertible were all mapped to the Car category. This mapping process introduces a potential challenge for the classifier-based method. The definitions for the Person class in ImageNet-1K, for instance, include categories like ballplayer and groom, which are not representative of typical pedestrians in a driving context and could potentially limit the classifier's performance on that specific class.

The final benchmark dataset consists of 1355 manually annotated instances distributed across a total of 400 images. To facilitate a balanced comparison of model performance across different conditions, an equal number of images (100) was selected for each of the four environmental scenarios described previously.

Figure 2 illustrates the distribution of these annotated instances across the five target classes. The distribution reveals a significant class imbalance, with car being the most frequent class (429 instances) and bus being the least frequent (75 instances). This imbalance is considered representative of real-world urban driving, where certain vehicle types are encountered more often than others. This

characteristic underscores the importance of using evaluation metrics, such as the macro F1-score, that are robust to imbalanced class distributions.

Table 1. Class label mapping.

COCO class ID	COCO class name	ImageNet-1K class ID	ImageNet-1K class name
0	Person	981	Ballplayer
		982	Groom
		983	Scuba diver
2	Car	436	Beach wagon
		468	Cab
		511	Convertible
		627	Limousine
		661	Model T
		705	Passenger car
		751	Racer
		817	Sports car
5	Bus	654	Minibus
		779	School bus
		874	Trolleybus
7	Truck	555	Fire engine
		569	Garbage truck
		675	Moving van
		717	Pickup
		864	Tow truck
		867	Trailer truck
9	Traffic light	920	Traffic light

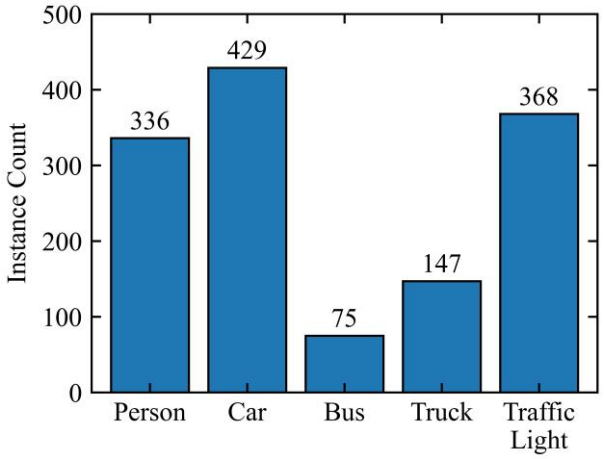
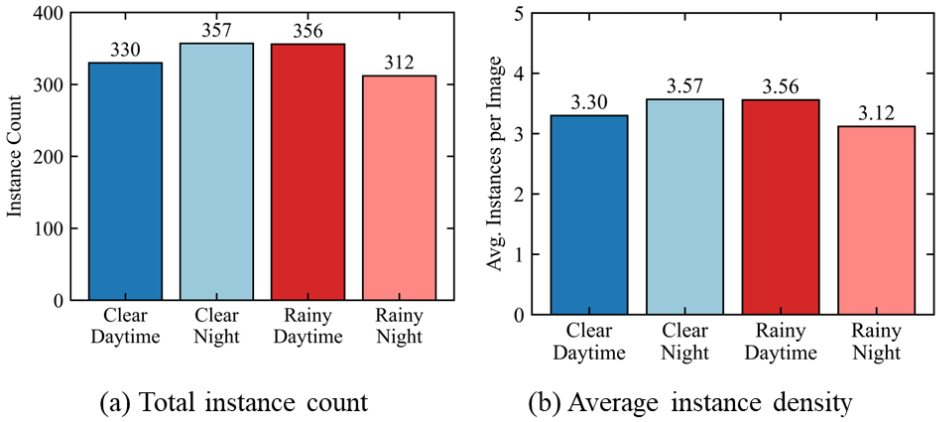


Figure 2. Distribution of annotated instances per class.

A more detailed analysis of the instance distribution and density across the four scenarios is provided in Figure 3. While the number of images per scenario is balanced, the number of annotated objects within them varies. Figure 3(a) shows the total instance count for each scenario. Figure 3(b) presents a more nuanced view by showing the average number of instances per image. This analysis reveals that the clear night and rainy daytime scenarios have the highest object density, with an average

of 3.57 and 3.56 instances per image, respectively. This suggests that these scenes are more visually cluttered, which could pose a greater perceptual challenge to the evaluated models.



(a) Total instance count (b) Average instance density

Figure 3. Distribution and density of instances across scenarios.

3. Assessment of paradigms

We evaluate and compare three vision-based paradigms, ranging from simple object detection approaches to those enhanced by segmentation-assisted object classification and VLM-based approaches.

3.1. Comparison of vision-based approaches

This section describes the three paradigms considered in this study and systematically compares five methods: one under Paradigm 1, two under Paradigm 2, and two under Paradigm 3.

1) Paradigm 1 (object detection–based): A direct identification approach utilizing the YOLOv13 model.
2) Paradigm 2 (segmentation-assisted classification): A two-stage pipeline evaluated through two specific methods: SAM2 combined with EfficientNetV2 and SAM2 combined with YOLOv13.
3) Paradigm 3 (VLM-based): A query-based approach assessed using two model scales (referred to as two methods): the 7-billion parameter version (Qwen2.5-VL-7b) and the 32B parameter version (Qwen2.5-VL-32b).

Figure 4 provides a high-level schematic that visually summarizes the core workflow of each of the three paradigms. The object detection–based paradigm, shown in Figure 4(a), is a single-step process where an object detection model processes the input image and the coordinate to directly output the class of the object containing the coordinate. The segmentation-assisted paradigm, depicted in Figure 4(b), is a two-stage pipeline. The first stage uses a segmentation model to isolate the object at the gaze coordinate, and the second stage uses a classification model to identify the resulting crop. The vision-language model paradigm, illustrated in Figure 4(c), is an end-to-end process where the input image, coordinate, and a structured text prompt are fed into a VLM, which directly outputs the semantic class.

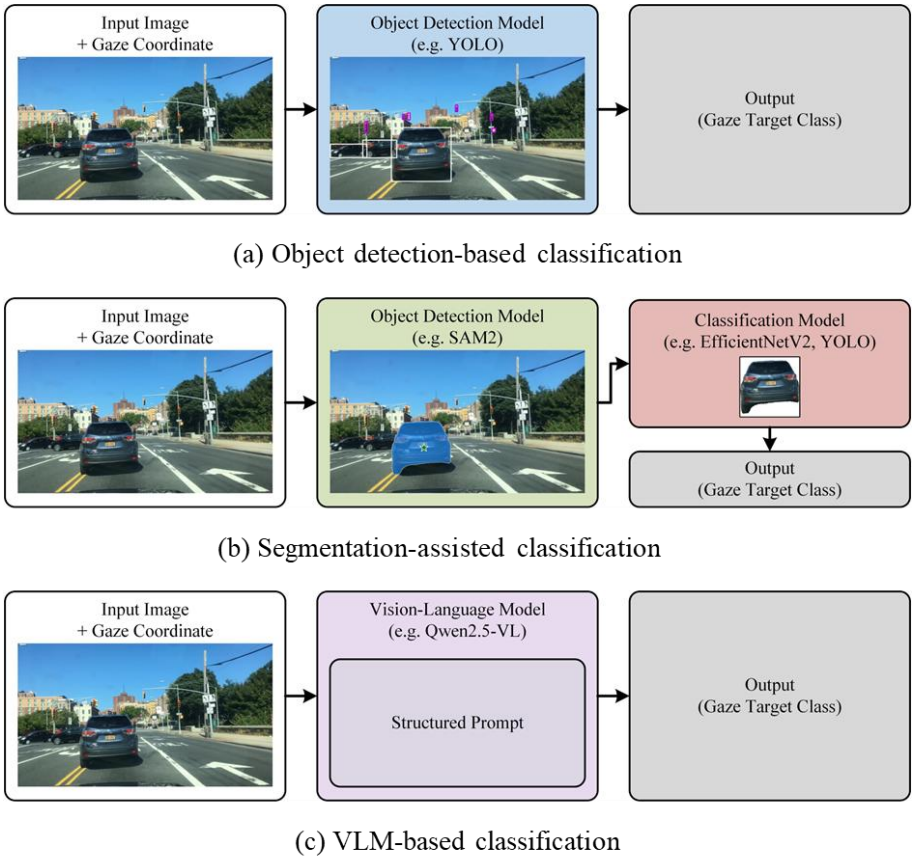


Figure 4. Visualization of the three methodological paradigms.

3.1.1. Paradigm 1: Object detection-based

This paradigm represents the most direct and computationally efficient solution. The specific implementation, Method 1, utilizes a pre-trained YOLOv13 model. The process begins by running the YOLOv13 model on the full input image, which generates a list of predicted bounding boxes, each associated with a class label and a confidence score. The decision logic then checks if the provided gaze coordinate falls within the geometric boundaries of a detected box. If it does, the class label of that box is assigned as the final prediction. A prediction of class ID -1 is recorded under three specific conditions: (1) the gaze coordinate does not fall within any detected bounding box, (2) the confidence score of the bounding box containing the coordinate is below a threshold of 0.4, or (3) the predicted class is not one of the five pre-defined target categories. This value indicates that the method did not identify a valid object at the designated point.

3.1.2. Paradigm 2: Segmentation-assisted classification

This paradigm aims to achieve higher precision by using pixel-level segmentation to isolate the object of interest prior to classification. The approach involves a two-stage pipeline common to two different methods. In the first stage, the SAM2 model is employed. The driver's gaze coordinate is provided to SAM2 as a positive point prompt, which guides the model to generate a precise pixel-level segmentation mask for the object at that location. Following segmentation, a standardized image preparation step is performed. The object is cropped from the original image based on the generated mask. To create a uniform input for the subsequent classification stage, a square image is constructed.

This new image places the cropped object at its center, and all background areas within the square are filled with a neutral white color. The two methods within this paradigm differ only in the second stage:

1) Method 2-1 (SAM2+EfficientNetV2): The prepared square image of the isolated object is passed to a pre-trained EfficientNetV2 model, which performs the final classification. If the class predicted by EfficientNetV2 does not correspond to one of the five target categories according to the mapping in Table 1, the final prediction of class ID is set to -1.
2) Method 2-2 (SAM2+YOLOv13): This method serves as a control experiment. The process is identical to Method 2-1, but the classification model is replaced with the YOLOv13 model. A prediction of class ID -1 is assigned if the class predicted by the YOLOv13 model is not one of the five target categories. This design allows for an analysis of whether performance limitations in this paradigm originate from the initial segmentation stage or the subsequent classification model.

3.1.3. Paradigm 3: VLM-based

The third paradigm explores the capabilities of large-scale VLMs to solve the task through contextual reasoning. This approach reframes the problem as a visual question answering task. The full input image, the (*x, y*) gaze coordinate, and a carefully structured text prompt are provided as a single, combined input to the VLM. The text prompt, detailed in Figure 5, is designed to explicitly ask the model to identify the object at the given coordinate. The prompt also constrains the possible answers to the predefined list of five target classes plus an "unsure" option. This "unsure" category effectively functions as the "-1" class, providing a designated output for cases of uncertainty and ensuring the model's responses are directly comparable to the other methods.

```
You are a precise image analysis assistant.

Your task is to identify the object located at a specific pixel coordinate within the provided image.


**Follow these strict rules:**

1. You must select the best-matching category from the following list: '[Person, Car, Bus, Truck, Traffic Light, Unsure]'.

2. Your response must be **strictly** one of the labels from the list (e.g., "Car"). **Do not** include any additional explanations, descriptions, or extraneous characters.

3. Prioritize the five specific object categories. Use "Unsure" only as the last resort when you are completely uncertain.


Now, analyze this image and tell me what object is located at the coordinate '[{x}, {y}]'.
```

Figure 5. The prompt used for the VLM-based methods.

Two methods were evaluated to investigate the impact of model scale on performance:

1) Method 3-1 (Qwen2.5-VL-7b) utilizes the 7-billion parameter version of the Qwen-VL model.

2) Method 3-2 (Qwen2.5-VL-32b) utilizes the much larger 32-billion parameter version of the model.

The comparison between these two models is intended to provide insight into how model scale affects performance on this specific and fine-grained visual identification task.

3.2. Evaluation metrics

To quantitatively assess the performance of the compared methodologies, a comprehensive set of standard classification metrics was adopted. The primary metrics used for evaluation were accuracy, precision, recall, and F1-score. Accuracy is defined as the ratio of all correct predictions to the total number of instances. For a more detailed analysis of error types, precision, recall, and F1-score were computed by Eqs (1)–(3), respectively.

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{2}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}, \tag{3}$$

where TP, FP, and FN represent the counts of true positives, false positives, and false negatives for a given class, respectively. Precision measures the model's ability to avoid making false positive predictions, while recall measures its ability to find all ground truth instances. The F1-score provides a single, balanced measure of performance by taking the harmonic mean of precision and recall.

A special consideration was made for handling model outputs that did not correspond to a target class. A prediction of class ID of "-1" signifies a failure by a model to assign one of the five target categories. Such failures can occur due to various method-specific conditions, including low confidence scores, the gaze point falling outside a valid detection region, or an out-of-scope classification. In the evaluation, any instance resulting in a "-1" prediction was treated as a false negative for its true ground truth class. This approach correctly penalizes a model's recall for its failure to identify a valid object, as the ground truth dataset contains labels for only the five target object classes.

Furthermore, to account for the significant class imbalance in the dataset, as established in Figure 2, macro-averaged metrics were used as the primary indicators of overall performance. Macro-precision, macro-recall, and the macro F1-score are calculated by computing the metric independently for each of the five target classes and then taking the unweighted average. This method gives equal weight to each class, regardless of its frequency, thereby providing a more fair and robust assessment of a model's ability to perform well across all object categories.

4. Results

This section presents a systematic evaluation of the five compared methodologies on the developed benchmark dataset. The analysis begins with an overview of the overall performance, followed by detailed investigations into model robustness across different scenarios and effectiveness on various object categories.

4.1. Overall performance comparison

First, the overall performance of the five vision-based methods was evaluated on the benchmark dataset. To ensure a comprehensive and fair assessment in the presence of inherent class imbalance, four evaluation metrics were considered: overall accuracy, macro-precision, macro-recall, and the macro F1-score.

Figure 6 presents the overall results of the study, comparing the evaluated methods in terms of overall accuracy and the more robust macro F1-score. As shown in Figure 6(a), the overall accuracy results clearly reveal the top two performers. The Qwen2.5-VL-32b model achieved the highest accuracy at 0.815, closely followed by the YOLOv13 model at 0.808. In contrast, the other three methods performed significantly worse, with the two SAM2-based approaches scoring below 0.201. Nonetheless, the overall accuracy can be misleading in the presence of class imbalance. A more suitable and robust metric for a class-imbalanced dataset is the macro F1-score, as plotted in Figure 6(b) . Using this metric, the YOLOv13 model (0.872) demonstrates a slightly better performance than the Qwen2.5-VL-32b model (0.845), while there is still a large performance gap for the other three methods.

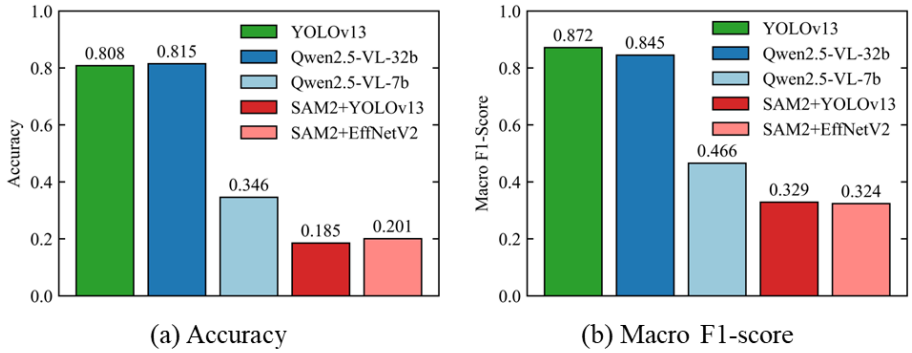
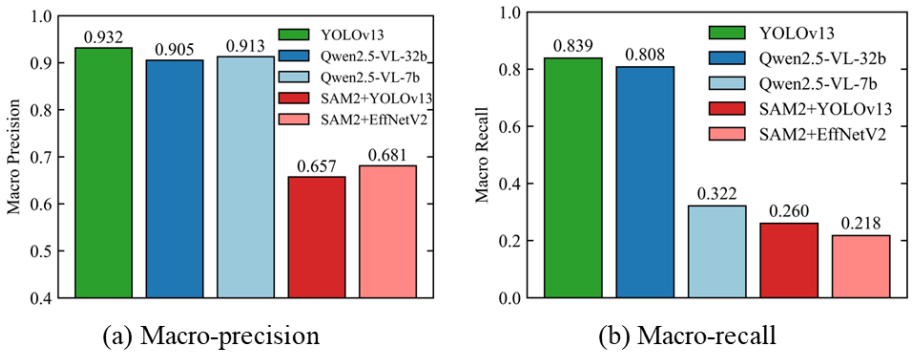


Figure 6. Overall performance comparison.

To gain a deeper understanding of the models' behavior, their performance was further assessed using macro-precision and macro-recall; the results are shown in Figure 7. Macro-precision quantifies the proportion of predicted positive instances that are correct. The analysis in Figure 7(a) reveals a crucial observation: the top three methods (YOLOv13 at 0.932, Qwen2.5-VL-32b at 0.905, and Qwen2.5-VL-7b at 0.913) all achieve very high macro-precision values above 0.90. This indicates that, across classes, when these models assign an object label, the prediction is highly likely to correspond to a true positive. In contrast, macro-recall measures the proportion of ground truth positive instances that are successfully identified, capturing a model's ability to comprehensively detect objects present in the scene. As shown in Figure 7(b), a clear difference in macro-recall is observed. While YOLOv13 (0.839) and Qwen2.5-VL-32b (0.808) maintain strong macro-recall, the other three methods exhibit very low macro-recall values, with the two SAM2-based pipelines achieving macro-recall values of only 0.260 and 0.218. This suggests that, despite producing relatively accurate predictions when detections occur, these methods fail to identify a substantial fraction of ground-truth objects.



(a) Macro-precision (b) Macro-recall

Figure 7. Performance comparison.

4.2. Performance analysis by scenario

To evaluate the robustness of these methods, their performance was analyzed across the four environmental scenarios. The overall accuracies are shown in Figure 8. A primary observation is that for nearly all methods, the highest performance is achieved under clear daytime conditions, while the rainy night condition proves to be the most challenging scenario. Across all four conditions, the YOLOv13 and Qwen2.5-VL-32b models demonstrate significantly better performance over the other methods. This accuracy comparison reveals a noteworthy pattern. Although YOLOv13 achieved a slightly higher accuracy in clear daytime (0.87 vs. 0.82), the Qwen2.5-VL-32b model is the top performer across the remaining three more challenging scenarios, particularly in nighttime settings (clear night: 0.82 vs. 0.73; rainy night: 0.78 vs. 0.75). This observation suggests that the VLM exhibits greater resilience under low-light conditions, warranting a deeper examination using more robust evaluation metrics.

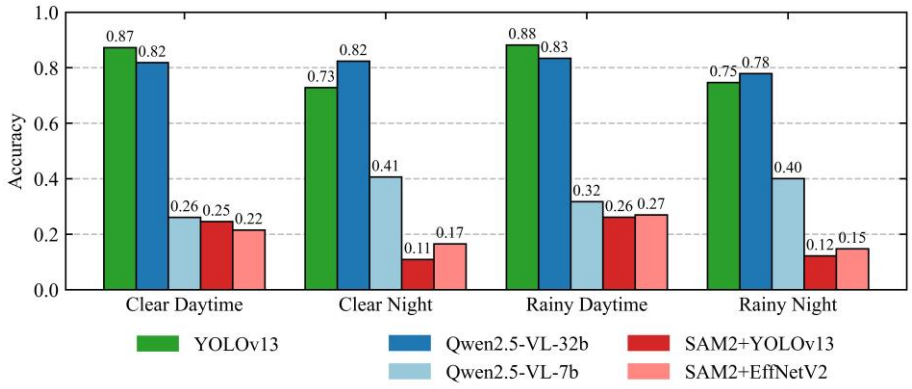


Figure 8. Comparison of accuracies across environmental scenarios.

A more rigorous analysis of model robustness was conducted using the macro F1-score, which accounts for class imbalance. As shown in Figure 9, the results corroborate the overall performance ranking and degradation trends observed with accuracy. However, a closer examination reveals that the YOLOv13 consistently achieves the highest macro F1-score across all four scenarios, indicating its better overall performance. For example, in the most challenging rainy night scenario, YOLOv13 attains a macro F1-score of 0.81, outperforming Qwen2.5-VL-32b, which achieves 0.78.

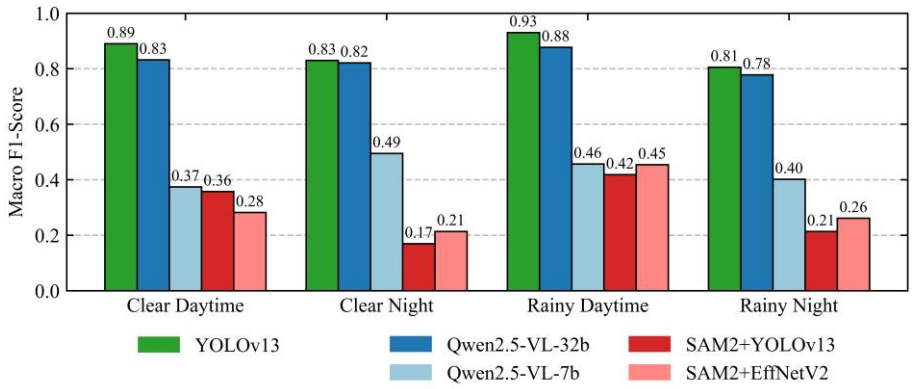


Figure 9. Comparison of macro F1-score across environmental scenarios.

Besides the macro F1-score, Figure 10 shows performance comparison in macro-precision and macro-recall. Figure 10(a) shows that macro-precision for the top-performing models (YOLOv13 and Qwen2.5-VL-32B) remains consistently high across all scenarios. In contrast, macro-recall exhibits a clear and systematic decline as environmental conditions worsen (Figure 10(b)). For example, YOLOv13's recall decreases from 0.88 in clear daytime to 0.76 in rainy night. This comparison reveals that performance degradation under adverse conditions is driven primarily by missed detections (false negatives), indicating that models increasingly fail to detect objects under challenging visual conditions.

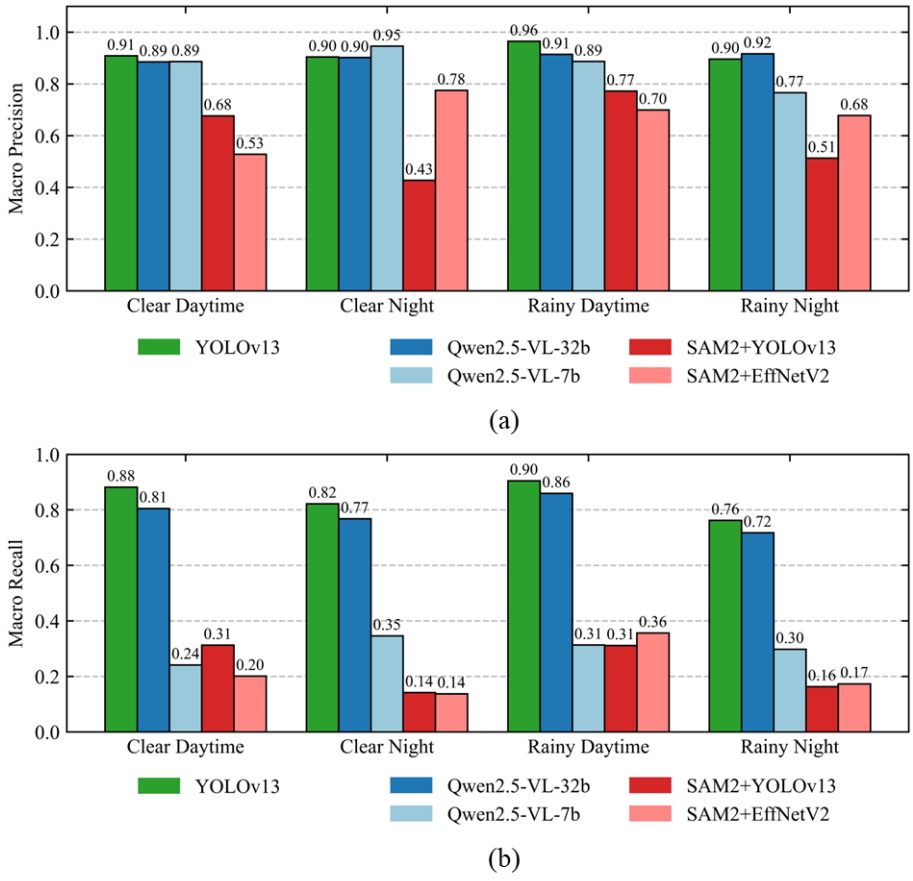


Figure 10. Performance comparison across environmental scenarios: (a) macro-precision; (b) macro-recall.

4.3. Performance analysis by object category

Following the analysis of performance across different environmental scenarios, this section examines the performance of each method across five object categories. The goal is to uncover paradigm-specific strengths and limitations with respect to different object types, such as large vehicles versus small signals and common versus rare classes. Similarly, the analysis begins with a general overall accuracy, followed by a more rigorous assessment using the F1-score, precision, and recall to provide a balanced and comprehensive evaluation.

Figure 11 presents overall accuracies across the five object categories. Again, YOLOv13 and Qwen2.5-VL-32b emerge as the top performers in most categories, while revealing complementary strengths. YOLOv13 excels on vehicle classes such as car (0.93) and bus (0.96), whereas Qwen2.5-VL-32b performs notably better on traffic light detection, achieving a much higher accuracy of 0.82 compared to YOLOv13's 0.57.

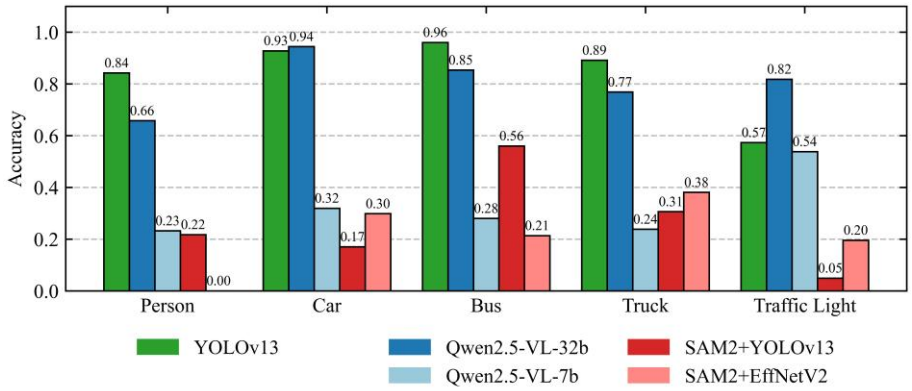


Figure 11. Performance comparison by accuracy across object categories.

Figure 12 shows the F1-score across the five object categories, confirming the strong performance of the YOLOv13 and Qwen2.5-VL-32b. YOLOv13 achieves the highest F1-scores on common, large objects—person (0.91), car (0.94), bus (0.91), and truck (0.87)—highlighting its strength in recognizing well-defined classes. Notably, a clear performance crossover emerges for the traffic light category: Qwen2.5-VL-32B attains a substantially higher F1-score (0.90) than YOLOv13 (0.73), revealing a distinct advantage of the large VLM paradigm for small and visually challenging objects.

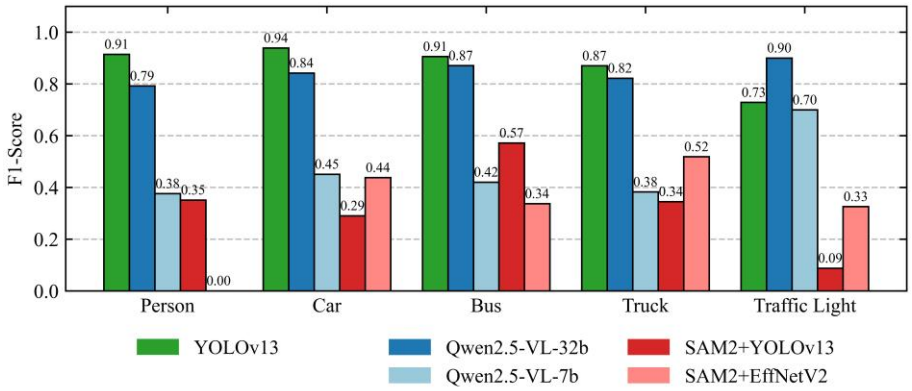


Figure 12. Performance comparison by F1-score across object categories.

Figure 13 shows class-level precision and recall. Both YOLOv13 and Qwen2.5-VL-32B achieve perfect precision (1.00) for the traffic light class, indicating no false positives. The performance gap is instead driven by recall: Qwen2.5-VL-32B attains a recall of 0.82, substantially higher than YOLOv13's 0.57. This explains YOLOv13's lower F1-score, which stems from missed detections rather than misclassification. VLM's superior recall likely reflects its ability to use a global scene context, enabling a more reliable detection of small, visually challenging objects.

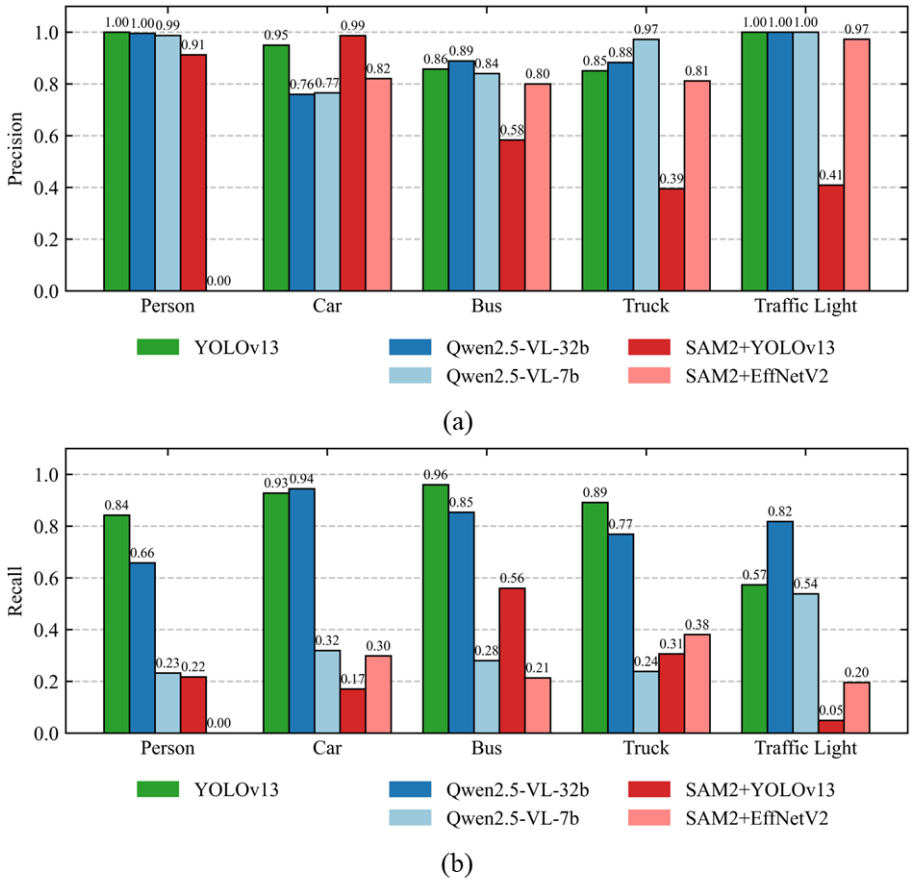


Figure 13. Performance comparison across object categories: (a) precision; (b) recall.

4.4. Analysis of model behavior

To complement the quantitative results, this section provides a deeper analysis of the internal behavior and qualitative error patterns of the evaluated models. This analysis proceeds in two parts: first, confusion matrices are used to visualize the class-specific error modes; second, confidence score distributions of key model components are examined to understand how the models arrive at their decisions and to diagnose the underlying causes of failures.

Figure 14 presents the confusion matrices for all five evaluated methods, allowing for a direct comparison of their error patterns. The matrices for the top-performing models, YOLOv13 (Figure 14(a)) and Qwen2.5-VL-32b (Figure 14(e)), exhibit strong diagonal dominance, visually confirming their high rate of correct classifications. For these models, the few off-diagonal errors represent logical misclassifications, such as the confusion between Truck (ID 7) and Car (ID 2), which is expected given their visual similarity. The most critical insight, however, comes from the first column of each matrix, which represents predictions of class ID -1. The matrices for both segmentation-assisted methods, SAM2+EfficientNetV2 (Figure 14(b)) and SAM2+YOLOv13

(Figure 14(c)), show extremely high values in this column. For instance, in Figure 14(b), 335 out of 336 person instances were incorrectly rejected. Similarly, the matrix for the smaller VLM, Qwen2.5-VL-7b (Figure 14(d)), also shows a very high number of rejections in this column compared to its larger counterpart. This provides direct visual evidence for the previously established finding that the catastrophic failure of the segmentation-assisted methods, and the comparatively weaker performance of the 7b VLM, is primarily due to an inability to make a positive classification (a recall problem), not a high rate of misclassification.

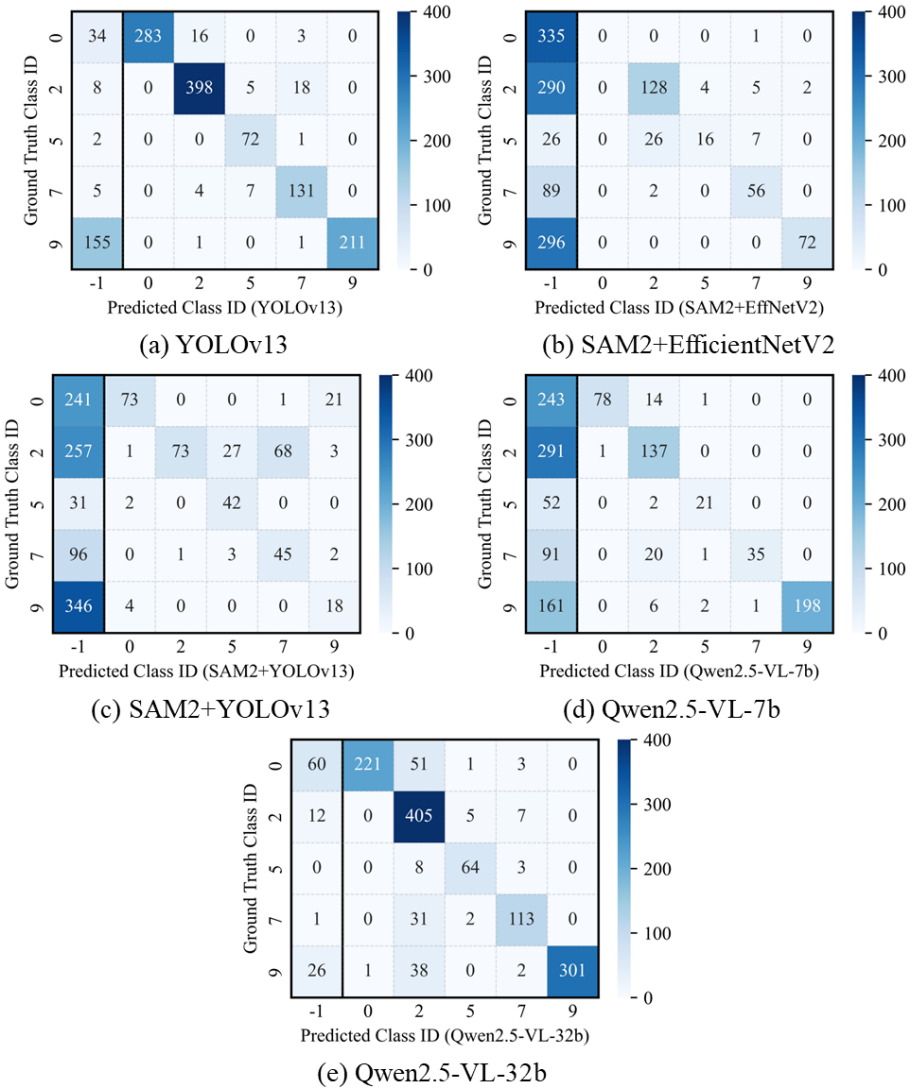


Figure 14. Confusion matrices for all five methods.

To further diagnose the underlying cause of the segmentation-assisted paradigm's failure, the confidence scores of its constituent components were analyzed, as shown in Figure 15. The histogram for the standalone YOLOv13 model in Figure 15(a) shows that it is generally very confident in its predictions, with a distribution heavily skewed toward high scores between 0.8 and 1.0. The key piece of evidence is presented in Figure 15(b), which shows that the SAM2 segmentation stage is also very confident, with its scores heavily concentrated above 0.8. However, when the classifiers are run on the crops generated by SAM2, their confidence decreases significantly. As seen in Figures 15(c) and (d),

the post-SAM2 classifiers exhibit much more uncertain and scattered confidence distributions, with the EfficientNetV2 model's confidence scores concentrated in the low 0.1–0.3 range.

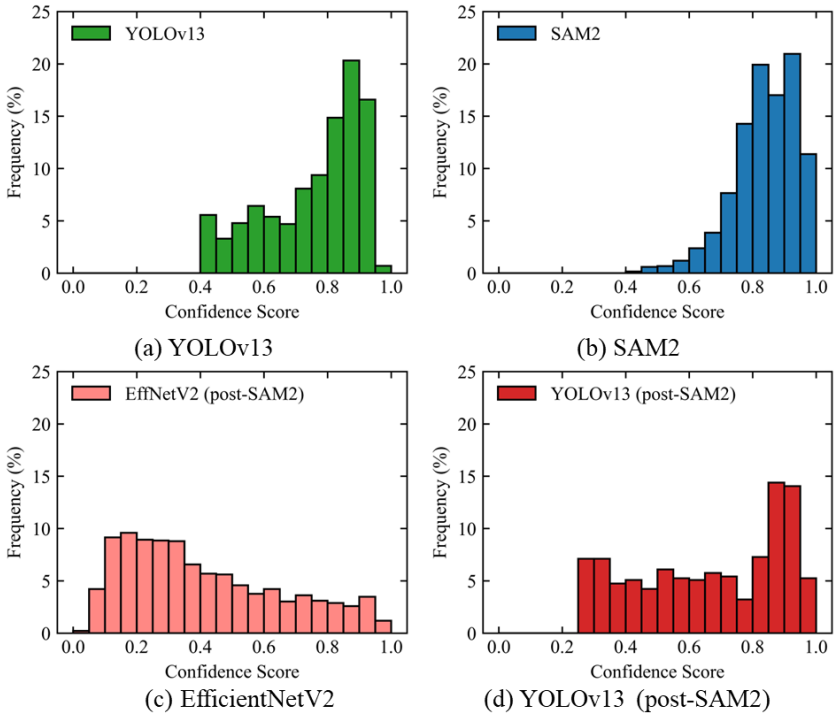


Figure 15. Confidence score distributions for model components.

These observations lead to a clear diagnosis of the *part-versus-whole* problem. The failure of the SAM2-based pipeline does not stem from poor or low-confidence segmentation; rather, SAM2 is often highly confident in its outputs. The issue arises because SAM2 confidently segments an incorrect semantic entity. For instance, when a driver's gaze falls on a car's taillight, SAM2 may accurately segment the taillight as an object part. The downstream classifier, however, is then presented with an isolated part and must select from whole-object categories such as "car", leading to uncertainty or misclassification. This semantic mismatch accounts for the low confidence scores and frequent background (Class ID: -1) predictions, identifying the *part-versus-whole* problem as a fundamental limitation of the segmentation-assisted paradigm for this task.

4.5. Analysis of computational footprint

Beyond model performance, computational efficiency is critical for real-world deployment in vehicles. To evaluate this, all experiments were conducted on a single NVIDIA A6000 GPU. Inference speed (FPS) was computed by averaging the processing time over 100 test images, while peak GPU memory usage reflects the maximum memory consumption observed during this process.

Table 2 summarizes the computational costs and resource requirements of the five methods, showing large differences in both speed and memory. YOLOv13 is the only model achieving real-time performance (36.1 FPS) with modest memory use (27.6M parameters, 1.53 GB VRAM), making it suitable for in-vehicle deployment. SAM2-based pipelines run at approximately 10 FPS (10.3–10.4 FPS), potentially acceptable for near real-time tasks but insufficient for latency-critical applications. The VLMs exhibit extremely low throughput, with Qwen2.5-VL-7B at 2.1 FPS and Qwen2.5-VL-32B

at 0.07 FPS (approximately 14 s per frame), alongside high memory demands (17.6–45 GB VRAM), necessitating a top-tier or customized GPU.

Combining these results with the previous performance analyses highlights a clear performance-efficiency trade-off. YOLOv13 achieves a strong balance of accuracy and speed, making it the most practical choice for real-time deployment. Qwen2.5-VL-32B, while robust and accurate in certain conditions, incurs prohibitive computational costs, limiting its real-time applicability. SAM2-based methods perform poorly in both efficiency and accuracy, making them the least viable option.

Table 2. Computational cost comparison.

Paradigm	Method	Model size (GB)	Parameters (M)	Inference speed (FPS)	Peak GPU memory usage (GB)
Object detection–based	YOLOv13	0.107	27.6	36.1	1.53
Segmentation-assisted classification	SAM2 + EfficientNetV2	0.963	343.0	10.4	11.83
	SAM2 + YOLOv13	1.3	252.1	10.3	9.20
VLM-based	Qwen2.5-VL-7b	15.5	8,292.2	2.1	17.63
	Qwen2.5-VL-32b	63.6	33,452.7	0.07	44.97

5. Discussion

The main finding of this study is that direct object detection and large-scale VLM paradigms outperform the segmentation-assisted classification approach for point-based object identification. This suggests that methods leveraging holistic scene context are more effective than those relying on an analytic, staged strategy. Both YOLOv13 and Qwen2.5-VL-32B, despite their different architectures, analyze the entire image to inform their predictions, enabling more robust inferences than the two-stage process of isolating a small, context-free region before classification.

The segmentation-assisted methods fail due to a fundamental *part-versus-whole* semantic gap. Confidence analysis (Figure 15) shows that SAM2 produces highly confident segmentations, while downstream classifiers exhibit low and unstable confidence, indicating that errors arise from confidently segmenting object parts rather than complete objects. For vehicles, gaze points often yield precise segmentations of components such as license plates, which are not classifiable as “car”. For small objects (person, traffic light), the segmented regions are frequently too small and context-poor for reliable recognition. Performance on “person” is further limited by a taxonomy mismatch in ImageNet-1K. In contrast, moderate success on larger objects (bus and truck) occurs when gaze points fall on broad object regions, allowing full-object segmentation and more reliable classification.

By contrast, the strong performance of the large VLM, particularly in challenging situations, highlights the advantage of its contextual and commonsense reasoning capabilities. The superiority of Qwen2.5-VL-32b was most evident in its robust performance for small objects, like “traffic light” (Figure 12), and its relative stability in adverse conditions (Figure 9). The precision-recall analysis for the “traffic light” class (Figure 13) is especially informative: despite perfect precision, the VLM’s substantially higher recall indicates its ability to leverage implicit world knowledge, effectively inferring that a small, colored light at an intersection is a traffic light. This advantage is most pronounced at night, where Qwen2.5-VL-32b significantly outperforms the traditional detector. Such holistic scene reasoning enables the VLM to remain effective even when local visual cues are degraded by low light or adverse weather.

These findings translate into clear practical implications for system engineers, summarized in Table 3. The results highlight a fundamental performance/efficiency trade-off. For current in-vehicle applications requiring real-time feedback (>10 FPS), an efficient object detector like YOLO represents the most practical choice, offering strong accuracy in common scenarios at minimal computational cost. However, for future safety-critical applications, such as Level 3+ ADAS or post-incident forensic analysis, where robustness under adverse and rare conditions is paramount, large VLMs represent a promising alternative. Although their substantial computational demands currently limit real-time applicability, their superior contextual reasoning and resilience in challenging environments point to the future of human-centric vehicle perception. Continued advances in hardware and model optimization are expected to narrow this gap.

Table 3. Summary of three paradigms and practical trade-offs.

Paradigm	Overall performance (macro F1)	Robustness (night/rain) (macro F1)	Small object performance	Speed (FPS)	VRAM (GB)	Primary disadvantage
Detection-based	High (0.87)	Stable (≥0.81)	Low (0.57)	Real-time (36.1)	Low (1.53)	Missed detections for small objects
Segmentation-assisted	Low (<0.35)	Vulnerable (≤0.45)	Minimal (≤0.20)	Near real-time (10.3)	High (9.20)	*Part-versus-whole* semantic gap
VLM-based	Scale-dependent (0.47–0.85)	Scale-dependent (0.40–0.88)	Scale-dependent (0.54–0.82)	Slow (0.07)	Very high (44.97)	High computational cost

6. Conclusions

This paper presents a systematic, cross-paradigm evaluation of methods for identifying the semantic object at a driver's point of gaze. The results lead to several key findings.

Direct object detection (YOLOv13) and large-scale vision-language model (Qwen2.5-VL-32b) paradigms were found to be the most effective approaches for the task. In contrast, the segmentation-assisted paradigm was shown to be fundamentally limited by a *part-versus-whole* semantic gap.

A clear trade-off was established between the two leading paradigms. YOLOv13 offers exceptional real-time efficiency and is the most practical choice for current in-vehicle deployment. Qwen2.5-VL-32b, while computationally demanding, provides better robustness and contextual understanding, particularly for small objects (e.g., traffic light) and under adverse conditions (e.g., nighttime), making it a promising candidate for future safety-critical applications.

Across all models, performance degradation in challenging conditions was driven primarily by missed detections (a recall problem). The large VLM (Qwen2.5-VL-32b) demonstrates consistently higher recall under these conditions, highlighting the advantage of its holistic and context-aware perception.

These findings provide guidance for the development of human-aware vehicle systems. Nonetheless, several limitations point to directions for future work. The benchmark dataset, while diverse, is limited in scale and relies on simulated gaze points that do not reflect the noise characteristics of real eye-tracking hardware. Furthermore, the current analysis is frame-based and does not exploit temporal context. Future research should therefore validate these findings using real-world gaze data, investigate hybrid architectures that balance efficiency and robustness, and extend VLM-based approaches to video inputs to leverage temporal information.

Use of AI tools declaration

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with language editing and improving grammar, clarity, and readability. The authors reviewed, revised, and verified all generated text and take full responsibility for the content of this publication.

Conflict of interest

Jidong J. Yang is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. *National Center for Statistics and Analysis, Early estimate of motor vehicle traffic fatalities in 2024*, National Highway Traffic Safety Administration, 2025. Available from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813710.
2. *WHO, Global Status Report on Road Safety 2023*, World Health Organization, 2023. Available from: https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023.
3. S. García-Herrero, J. Febres, W. Boulagouas, J. Gutiérrez, M. Saldaña, Assessment of the influence of technology-based distracted driving on drivers' infractions and their subsequent impact on traffic accidents severity, *Int. J. Environ. Res. Public Health*, **18** (2021), 7155. https://doi.org/10.3390/ijerph18137155
4. C. Ahlström, K. Kircher, M. Nyström, B. Wolfe, Eye tracking in driver attention research—how gaze data interpretations influence what we learn, *Front. Neuroergonomics*, **2** (2021), 778043. https://doi.org/10.3389/fnrgo.2021.778043
5. *National Center for Statistics and Analysis, Distracted driving in 2023*, National Highway Traffic Safety Administration, 2025. Available from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813703.
6. S. Klauer, J. Ehsani, D. McGehee, M. Manser, The effect of secondary task engagement on adolescents' driving performance and crash risk, *J. Adolescent Health*, **57** (2015), S36–S43. https://doi.org/10.1016/j.jadohealth.2015.03.014
7. B. Simons-Morton, F. Guo, S. Klauer, J. Ehsani, A. Pradhan, Keep your eyes on the road: young driver crash risk increases according to duration of distraction, *J. Adolescent Health*, **54** (2014), S61–S67. https://doi.org/10.1016/j.jadohealth.2013.11.021
8. F. Guo, S. Klauer, Y. Fang, J. Hankey, J. Antin, M. Perez, et al., The effects of age on crash risk associated with driver distraction, *Int. J. Epidemiol.*, **46** (2017), 258–265. https://doi.org/10.1093/ije/dyw234
9. Y. Dong, Z. Hu, K. Uchimura, N. Murayama, Driver inattention monitoring system for intelligent vehicles: a review, *IEEE Trans. Intell. Transp.*, **12** (2011), 596–614. https://doi.org/10.1109/TITS.2010.2092770
10. P. Deng, C. Cui, Z. Cheng, Q. Zhang, Y. Bu, Fatigue damage prognosis of orthotropic steel deck based on data-driven LSTM, *J. Constr. Steel Res.*, **202** (2023), 107777. https://doi.org/10.1016/j.jcsr.2023.107777
11. P. Deng, J. Yang, T. Yee, Deep learning-based flood detection for bridge monitoring using accelerometer data, *Infrastructures*, **9** (2024), 140. https://doi.org/10.3390/infrastructures9090140
12. M. Khan, S. Lee, Gaze and eye tracking: techniques and applications in ADAS, *Sensors*, **19** (2019), 5540. https://doi.org/10.3390/s19245540

13. A. Ledezma, V. Zamora, O. Sipele, M. Sesmero, A. Sanchis, Implementing a gaze tracking algorithm for improving advanced driver assistance systems, *Electronics*, **10** (2021), 1480. https://doi.org/10.3390/electronics10121480
14. Y. Feng, Y. Chen, J. Zhang, C. Tian, R. Ren, T. Han, et al., Human-centred design of next generation transportation infrastructure with connected and automated vehicles: a system-of-systems perspective, *Theor. Iss. Ergon. Sci.*, **25** (2024), 287–315. https://doi.org/10.1080/1463922X.2023.2182003
15. J. Kim, Sustainable real-time driver gaze monitoring for enhancing autonomous vehicle safety, *Sustainability*, **17** (2025), 4114. https://doi.org/10.3390/su17094114
16. S. Haghzare, J. Campos, A. Mihailidis, Classifying older drivers' gaze behaviour during automated versus non-automated driving: a preliminary step towards detecting mode confusion, *Int. J. Hum.-Comput. Int.*, **40** (2024), 241–254. https://doi.org/10.1080/10447318.2022.2112933
17. F. Walker, J. Wang, M. Martens, W. Verwey, Gaze behaviour and electrodermal activity: objective measures of drivers' trust in automated vehicles, *Transport. Res. F-Traf.*, **64** (2019), 401–412. https://doi.org/10.1016/j.trf.2019.05.021
18. Y. Zhu, Y. Geng, R. Huang, X. Zhang, L. Wang, W. Liu, Driving towards the future: exploring human-centered design and experiment of glazing projection display systems for autonomous vehicles, *Int. J. Hum.-Comput. Int.*, **40** (2024), 4087–4102. https://doi.org/10.1080/10447318.2023.2209836
19. M. Winlaw, S. Steiner, R. MacKay, A. Hilal, Using telematics data to find risky driver behaviour, *Accident Anal. Prev.*, **131** (2019), 131–136. https://doi.org/10.1016/j.aap.2019.06.003
20. D. Kumar, N. Muhammad, Object detection in adverse weather for autonomous driving through data merging and YOLOv8, *Sensors*, **23** (2023), 8471. https://doi.org/10.3390/s23208471
21. G. Liu, K. Wu, W. Lan, Y. Wu, YOLO-FDCL: improved YOLOv8 for driver fatigue detection in complex lighting conditions, *Sensors*, **25** (2025), 4832. https://doi.org/10.3390/s25154832
22. T. Tammi, J. Pekkanen, B. Cowley, O. Lappi, Quantifying tracking quality during occlusion with an integrated gaze metric anchored to task performance, *Sci. Rep.*, **15** (2025), 31858. https://doi.org/10.1038/s41598-025-17519-8
23. D. Mardanbegi, T. Langlotz, H. Gellersen, Resolving target ambiguity in 3d gaze interaction through vor depth estimation, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, 1–12. https://doi.org/10.1145/3290605.3300842
24. J. Wolfe, A. Kosovicheva, B. Wolfe, Normal blindness: when we look but fail to see, *Trends Cogn. Sci.*, **26** (2022), 809–819. https://doi.org/10.1016/j.tics.2022.06.006
25. M. Herslund, N. Jørgensen, Looked-but-failed-to-see-errors in traffic, *Accident Anal. Prev.*, **35** (2003), 885–891. https://doi.org/10.1016/S0001-4575(02)00095-7
26. K. Kennedy, J. Bliss, Inattentional blindness in a simulated driving task, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **57** (2013), 1899–1903. https://doi.org/10.1177/1541931213571423
27. J. Hoffmann, H. Tosso, M. Santos, J. Justo, A. Malik, A. Rahman, Real-time adaptive object detection and tracking for autonomous vehicles, *IEEE Transactions on Intelligent Vehicles*, **6** (2021), 450–459. https://doi.org/10.1109/TIV.2020.3037928
28. Z. Wang, K. Yao, F. Guo, Driver attention detection based on improved YOLOv5, *Appl. Sci.*, **13** (2023), 6645. https://doi.org/10.3390/app13116645
29. N. Youssouf, Traffic sign classification using CNN and detection using faster-RCNN and YOLOV4, *Heliyon*, **8** (2022), e11792. https://doi.org/10.1016/j.heliyon.2022.e11792
30. M. Maity, S. Banerjee, S. Chaudhuri, Faster R-CNN and yolo based vehicle detection: a survey, *Proceedings of the 5th International Conference on Computing Methodologies and Communication*, 2021, 1442–1447. https://doi.org/10.1109/ICCMC51019.2021.9418274

31. F. Tonini, N. Dall'Asen, C. Beyan, E. Ricci, Object-aware gaze target detection, *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, 21803–21812. https://doi.org/10.1109/ICCV51070.2023.01998
32. B. Mirzaei, H. Nezamabadi-Pour, A. Raoof, R. Derakhshani, Small object detection and tracking: a comprehensive review, *Sensors*, **23** (2023), 6887. https://doi.org/10.3390/s23156887
33. H. Zhao, S. Wang, X. Peng, J. Pan, R. Wang, X. Liu, Road surface semantic segmentation for autonomous driving, *PeerJ Comput. Sci.*, **10** (2024), e2250. https://doi.org/10.7717/peerj-cs.2250
34. F. Hazzaa, I. Udoidiong, A. Qashou, S. Yousef, Segment anything: a review, *Mesopotamian Journal of Computer Science*, **2024** (2024), 150–161. https://doi.org/10.58496/MJCSC/2024/012
35. Z. Yuan, Principles, applications, and advancements of the segment anything model, *Appl. Comput. Eng.*, **53** (2024), 73–78. https://doi.org/10.54254/2755-2721/53/20241270
36. N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, et al., Sam 2: segment anything in images and videos, *Proceedings of International Conference on Learning Representations (ICLR)*, 2025, 1–44.
37. M. Elhenawy, H. Ashqar, A. Rakotonirainy, T. Alhadidi, A. Jaber, M. Tami, Vision-language models for autonomous driving: clip-based dynamic scene understanding, *Electronics*, **14** (2025), 1282. https://doi.org/10.3390/electronics14071282
38. A. Keskar, S. Perisetla, R. Greer, Evaluating multimodal vision-language model prompting strategies for visual question answering in road scene understanding, *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2025, 937–946. https://doi.org/10.1109/WACVW65960.2025.00115
39. J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, et al., Qwen-vl: a versatile vision-language model for understanding, localization, text reading, and beyond, arXiv: 2308.12966. https://doi.org/10.48550/arXiv.2308.12966
40. S. Wang, D. Kim, A. Taalimi, C. Sun, W. Kuo, Learning visual grounding from generative vision and language model, *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, 8057–8067. https://doi.org/10.1109/WACV61041.2025.00782
41. F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, et al., Bdd100k: a diverse driving dataset for heterogeneous multitask learning, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 2636–2645. https://doi.org/10.1109/CVPR42600.2020.00271
42. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft coco: common objects in context, In: *Computer Vision-ECCV 2014*, Cham: Springer, 2014, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
43. M. Lei, S. Li, Y. Wu, H. Hu, Y. Zhou, X. Zheng, et al., Yolov13: real-time object detection with hypergraph-enhanced adaptive visual perception, arXiv: 2506.17733. https://doi.org/10.48550/arXiv.2506.17733
44. M. Tan, Q. Le, Efficientnetv2: smaller models and faster training, *Proceedings of the 38th International Conference on Machine Learning*, 2021, 10096–10106.

AIMS Press