Toward Inherently Robust VLMs Against Visual Perception Attacks

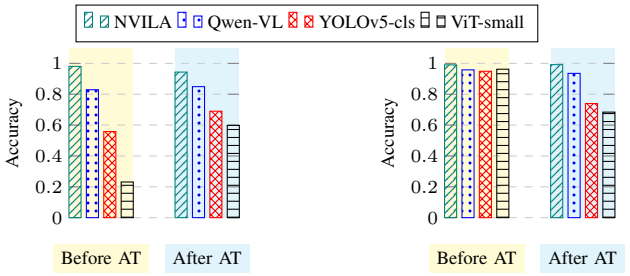
Pedram MohajerAnsari[1], Amir Salarpour[1], Michael Kühr[2], Siyu Huang[1], Mohammad Hamad[2], Habeeb Olufowobi[3], Sebastian Steinhorst[2], Bing Li[1], Mert D. Pesé[1]

***Abstract*— Autonomous vehicles rely on deep neural networks (DNNs) for traffic sign recognition, lane centering, and vehicle detection, yet these models are vulnerable to attacks that induce misclassification and threaten safety. Existing defenses (e.g., adversarial training) often fail to generalize and degrade clean accuracy. We introduce Vehicle Vision–Language Models (*V^2LMs*), fine-tuned Vision Language Models (VLMs) specialized for AV perception, and show that they are inherently more robust to unseen attacks without adversarial training, maintaining substantially higher adversarial accuracy than conventional DNNs. We study two deployments: *Solo* (task-specific *V^2LMs*) and *Tandem* (a single *V^2LM* for all three tasks). Under attacks, DNNs drop 33%–74%, whereas *V^2LMs* decline by under 8% on average. Tandem achieves comparable robustness to Solo while being more memory-efficient. We also explore integrating *V^2LMs* in parallel with existing perception stacks to enhance resilience. Our results suggest *V^2LMs* are a promising path toward secure, robust AV perception. Code and data are available at `https://github.com/pedram-mohajer/V2LM`.**

I. Introduction

Autonomous vehicles (AVs) rely on automated driving systems (ADS) consisting of perception, planning, and control. The perception module uses deep neural network (DNN) models for image classification and object detection, processes camera and LiDAR data [1], and passes the resulting scene understanding to the planning module for navigation decisions, which are then executed by the control module [2]. Prior point-cloud analysis methods include plane-kernel convolutions [3] and non-parametric designs using Gaussian positional encoding [**?**]. Although these DNN-based perception models recognize, monitor, and predict nearby objects [4], they are vulnerable to physical attacks that manipulate real-world environments to deceive perception [5], causing misclassification and unsafe driving behaviors, as shown by the DRP-Attack [6] on lane detection. Similar reliability concerns also arise across other applied ML pipelines beyond autonomous driving, including medical imaging and analytics and scientific simulation workflows [7], [8], [9].

Various defense methods have been proposed to mitigate such attacks: defensive distillation reduces model sensitivity but struggles to generalize across diverse perturbations [10]; input transformations can suppress adversarial noise but often degrade clean data quality, lowering accuracy [11]; and provable defenses, though theoretically robust, are computationally expensive and difficult to scale [12]. Among these, adversarial training is the most widely adopted defense [13].



(a) Accuracy on **adversarial** (b) Accuracy on **benign**

Fig. 1. Impact of adversarial training (AT) on NVILA, Qwen-VL, YOLOv5-cls, and ViT-small. (a) Adversarial inputs. (b) Benign inputs.

To address these limitations and achieve robust and generalizable models for AV perception tasks, this work proposes a novel finding: *VLMs inherently exhibit robustness against unseen adversarial attacks on AV perception systems even without adversarial training, significantly outperforming DNNs.* VLMs such as `Qwen-VL` [14] and `NVILA` [15] demonstrate strong resilience against adversarial inputs without adversarial training. Based on this observation, we are the first to systematically fine-tune VLMs for AV perception tasks; fine-tuning trains these VLMs on TSR, ALC, and VD, enabling a fair comparison with task-specific DNNs, and we name them **V**ehicle **V**ision **L**anguage **M**odels (*V^2LMs*).

Thus, we benchmark both conventional DNNs and *V^2LMs before and after* adversarial training (AT) on clean and adversarial inputs under ***unseen*** adversarial attacks ("unseen" means these attacks are never used during training or fine-tuning of the DNNs or VLMs and appear only at test time). As shown in Figure 1, this comparison highlights two findings. **(i) VLMs exhibit robustness:** before AT, `NVILA-8B` remains near 98% and `Qwen-VL` about 83% on adversarial inputs, while both stay above 95% on benign inputs; after AT, these numbers change by no more than ± 3 percentage points (pp), underscoring their insensitivity to the usual AT trade-offs [16]. **(ii) Conventional DNNs collapse under attack, and AT provides only modest relief—and at a cost:** `YOLOv5-cls` [17] plunges from roughly 95% to 56% (-39 pp), and `ViT-small` [18] drops from 96% to 23% (-73 pp); although AT increases adversarial accuracy to 69% for `YOLOv5-cls` and 60% for `ViT-small`, their benign

[1]Pedram MohajerAnsari, Amir Salarpour, Siyu Huang, Bing Li, and Mert D. Pesé are with Clemson University, Clemson, SC, USA {`pmohaje`, `asalarp`, `siyuh`, `bli4`, `mpese`}@clemson.edu

[2]Michael Kühr, Mohammad Hamad, and Sebastian Steinhorst are with the Technical Universität München, Munich, Germany {`michael.kuehr`, `mohammad.hamad`, `sebastian.steinhorst`}@tum.de

[3]Habeeb Olufowobi is with the University of Texas at Arlington, Arlington, TX, USA `habeeb.olufowobi@uta.edu`

TABLE I

CLASSIFICATION ACCURACY (%) OF YOLOV5-CLS AND VIT-SMALL ON GTSRB WITHOUT DEFENSES (ROW **BASE MODEL**) AND ACCURACY CHANGE Δ (PP) WHEN STANDARD DEFENSES ARE APPLIED TO BENIGN AND SHADOW IMAGES. GREEN (↑) AND RED (↓) ENTRIES DENOTE IMPROVEMENT OR DEGRADATION VS. THE NO-DEFENSE **BASE MODEL**.

	YOLOv5-cls		ViT-small	
Method	**benign**	**shadow**	**benign**	**shadow**
Base Model	94.88	55.82	96.20	23.12
Label Smoothing (LS)	↑ 0.87	↓ 3.71	↓ 2.47	↓ 11.71
Dropout (DO)	↓ 3.17	↓ 8.26	↓ 1.85	↓ 10.06
JPEG Compression (JPG)	↓ 2.81	↓ 11.79	↓ 1.74	↓ 8.03
Bit-Depth Reduction (BD)	↓ 1.75	↓ 0.95	↓ 0.53	↑ 2.27
Random Resized Cropping (RRP)	↓ 22.00	↓ 25.42	↓ 4.22	↓ 14.53
Histogram Equalization (HEQ)	↑ 1.59	↓ 0.82	↓ 1.03	↑ 1.12

accuracies simultaneously sink to 74% and 68%, reflecting the clean-data penalty noted in previous work [19].

We evaluate six VLMs (`LLaVA-7B`, `MobileVLM` [20], `LLaVA-13B-LoRA` [21], `MoE-LLaVA` [22], `Qwen-VL-7B` [14], and `NVILA-8B` [15]) as auxiliary perception modules for traffic-sign recognition (TSR), automated lane centering (ALC), and vehicle detection (VD) tasks. Each model is tested zero-shot and then re-evaluated after task-specific fine-tuning (**RQ1**). The first four models are LLaMA-based, the last two Qwen-based, chosen for their performance and offline suitability. Task-specific DNN baselines such as `YOLOv5-cls` and `ViT-small` for TSR, `CLRerNet` for ALC, and `YOLOv5-dt` for VD are fine-tuned on the same data to ensure a fair comparison with the VLMs under both benign and adversarial conditions.

Then, the effectiveness and resilience of both DNN models and *V²LMs* are tested using *unseen* (not used during model training or fine-tuning) adversarial examples (AEs) against three distinct attacks targeting AV perception algorithms (**RQ2**): (1) Robust and Accurate UV-map-based Camouflage attack (RAUCA) to deceive VD algorithms [23], (2) a physical-world adversarial attack known as the Dirty Road Patch (DRP-Attack) to compromise DNN-based automated lane centering (ALC) models, and (3) shadows cast on traffic signs to attack TSR algorithms [24]; these are strong physical attacks evaluated end-to-end and shown to cause safety-relevant failures.

The study then compares two distinct designs for utilizing *V²LMs* and evaluates their performance on the three aforementioned AV tasks and AEs (**RQ3**). The first design, termed *Solo Mode*, involves separate *V²LMs*, each fine-tuned individually for one of the AV tasks. The second design, named *Tandem Mode*, uses a single *V²LM* fine-tuned simultaneously for all three AV tasks, aiming to assess whether it can match the robustness of *Solo Mode* across tasks under both benign and adversarial conditions. This paper makes the following contributions:

- This work presents a novel finding: fine-tuned VLMs inherently exhibit superior robustness against ***unseen*** adversarial attacks compared to task-specific DNNs. We highlight a critical limitation of adversarial training: it significantly degrades benign accuracy while providing only limited improvements against adversarial examples. In contrast, VLMs achieve strong adversarial robustness while maintaining high benign accuracy.
- We introduce Vehicle Vision Language Models (*V²LMs*) which are fine-tuned VLMs for AV perception tasks: TSR, ALC, and VD. We propose two deployment strategies: *Solo Mode* (separate VLM fine-tuned for each task) and *Tandem Mode* (a single unified *V²LM* across multiple perception tasks).
- We conduct comprehensive experiments to evaluate the robustness of DNN models and *V²LMs* under adversarial conditions. DNN models experience performance drops of 33%–74% under attacks, whereas *V²LMs* show reductions of less than 8% on average, maintaining high adversarial accuracy without additional defense mechanisms.

II. RELATED WORK

Large Language Models in Autonomous Driving. Recent works have demonstrated the potential of LLMs in the context of AVs, particularly enhancing perception, control, and motion planning tasks. Regarding perception systems, LLMs utilize external APIs to access real-time information sources, including traffic reports, and weather updates, which significantly enrich the vehicle's ability to gain a comprehensive understanding of its environment [25]. Aldeen *et al.* [26] investigate the application of Large Multimodal Models (LMMs) for enhancing the cybersecurity of AVs. Concerning control, LLMs enable the adjustment of control settings according to driver preferences, thereby personalizing the driving experience [27].

Adversarial Threats and Defenses. Adversarial Examples (AEs) were first defined by Szegedy *et al.* [28] as subtly modified inputs designed to fool DNNs. These minor, often imperceptible alterations can drastically alter a DNN's predictions [29], [30]. To counter these vulnerabilities, various defense mechanisms have been proposed. We re-implemented these methods and evaluated them against the shadow attack [24]; as summarized in Table I, they are largely ineffective and often reduce benign accuracy. Label Smoothing (LS) [31] softens targets and can underfit fine-grained TSR cues. Dropout (DO) [32] removes capacity needed for small text and borders. JPEG Compression (JPG) [11] suppresses high-frequency edges and characters. Bit-Depth Reduction (BD) [33] distorts class-specific hues via quantization. Random Resized Cropping (RRP) [34] can crop out discriminative regions and destabilize scale. Histogram Equalization (HEQ) [35] over-equalizes, amplifying halos and shifting the distribution.

(a) Shadow Attack (b) DRP-Attack (c) RAUCA

Fig. 2. Examples of adversarial attacks targeting AV perception.

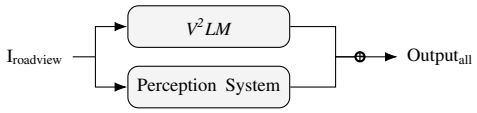


Fig. 3. Possible integration of V^2LM with perception system for attack mitigation. $I_{roadview}$ is the image captured by AV cameras, with the output sent to the planning and control modules for action.

III. V^2LM as a Defense Mechanism

A. Autonomous Vehicle Perception

The perception system in AVs is responsible for interpreting the environment through sensor data, such as images, to enable safe and efficient operation. This system performs essential tasks including TSR, ALC, and and VD which is part of object detection. TSR enables the vehicle to follow road rules by recognizing traffic signs [36], ALC keeps the vehicle centered within its lane by identifying road markings [37], and VD detects and classifies other vehicles [38]. The perception system (PS) can be represented as a function, shown in Equation 1, which takes an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and outputs the results of TSR, ALC, and VD. Specifically, TSR outputs the detected traffic sign's class and bounding box if a sign is present; ALC provides a classification for the appropriate steering command; and VD returns the bounding box and class of any detected car.

$$PS(I) = \begin{cases} TSR(\mathcal{I}) : Cls_{traffic\text{-}sign} \\ ALC(\mathcal{I}) : Cls_{steering} \\ VD(\mathcal{I}) : (Cls_{vehicle}, BBox_{vehicle}) \end{cases} \quad (1)$$

where *Cls* denotes class and *BBox* denotes bounding box.

AEs can interfere with the perception system by causing errors in these modules. An AE, $\mathcal{I}_{\text{adv}}$, is crafted by adding a small perturbation δ to an original input $\mathcal{I}$ such that the target module's output changes undesirably:

$$\mathcal{I}_{adv} = \mathcal{I} + \delta, \quad where \quad \|\delta\| \leq \epsilon \quad and \quad f(\mathcal{I}_{adv}) \neq f(\mathcal{I}) \quad (2)$$

where f denotes the specific target module in the AV's perception system, which can be the TSR, ALC, or VD module. This perturbation δ, constrained by ϵ, represents a small, controlled modification designed to be imperceptible, thereby ensuring that $\mathcal{I}_{\text{adv}}$ visually resembles $\mathcal{I}$.

B. Vision Language Model (VLM)

Definition. VLMs, by integrating visual and textual data [39], have the potential to enhance perception tasks by enabling a more comprehensive understanding of the environment. A pre-trained VLM takes an image and a text prompt as inputs to generate a relevant response:

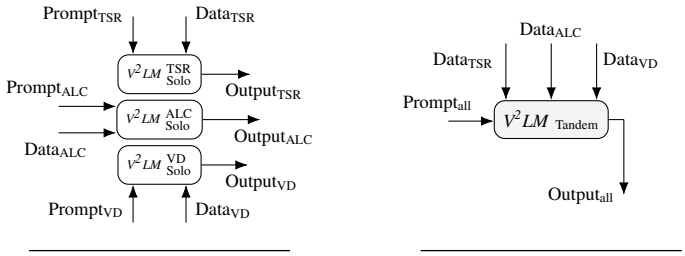
$$VLM_{pre}(\mathcal{I}, Prompt) \rightarrow Generated\text{-}Response \quad (3)$$

This potential arises from their several key strengths: their multimodal learning capability, which allows them to correlate visual and textual information simultaneously [40]; their robustness to variability, which enables them to generalize well across different environments due to extensive training on diverse datasets [41]; their contextual understanding, which leverages textual data to enhance the interpretation of visual scenes [42]; and their comprehensive feature extraction, which combines features from both visual and textual data [43]. For instance, VLMs could enhance ALC by interpreting road markings and reading associated signs. In VD, VLMs could recognize and classify objects like vehicles by analyzing visual data along with bounding box coordinates. Similarly, VLMs could improve TSR by understanding text on signs, such as speed limits or warnings, to ensure the car follows road rules accurately.

Fine-Tuning. Fine-tuning VLMs on AV-specific tasks is essential for optimizing their performance and enabling adaptation to domain-specific challenges such as variations in lighting, road conditions, and traffic scenarios; the resulting fine-tuned VLM is referred to as V^2LM. For the LLaVA family of models, this fine-tuning process involves adjustments to key components. The vision encoder, implemented as CLIP ViT-L/14 [44], extracts visual features from input images. The language model, based on Vicuna (a fine-tuned variant of LLaMA) [45], interprets textual prompts. A cross-modal attention module integrates the modalities, and all components are jointly optimized via visual instruction tuning [46] to align with autonomous vehicle (AV) tasks.

Qwen-VL adopts a modular architecture comprising an OpenCLIP ViT-bigG visual encoder [47], a single-layer cross-attention adapter with learnable queries, and a Qwen-7B language model [48]. The adapter compresses image features to fixed-length sequences, enabling efficient visual grounding and fine-grained perception. Its three-stage training pipeline includes weakly supervised pretraining on large-scale image-text pairs, multi-task visual-language pretraining, and supervised instruction tuning. NVILA builds on this by incorporating a "scale-then-compress" design, which first increases spatial and temporal input resolution and then compresses visual tokens for efficient processing. Its architecture consists of a SigLIP-based vision encoder [49], a lightweight MLP projector, and a Qwen2-based language model [50]. For fine-tuning, NVILA applies lower learning rates to the vision encoder's LayerNorms while optimizing the language model. This allows robust AV adaptation under limited compute budgets [15].

LoRA (Low-Rank Adaptation) [51] offers an efficient



(a) Solo Mode (b) Tandem Mode

Fig. 4. Comparison of Solo and Tandem Modes.

TABLE II

RANGE OF ZERO-SHOT PERFORMANCE (%) ACROSS THESE SIX VLMS FOR EACH AV PERCEPTION TASK. RESULTS REFLECT THE LOWEST AND HIGHEST VALUES AMONG ALL MODELS BEFORE FINE-TUNING.

Task	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
TSR	1.19–31.24	0.95–29.80	1.15–29.24	0.48–30.31
ALC	20.71–50.91	24.60–51.18	28.38–51.44	21.39–50.91
VD	60.64–92.06	64.01–92.01	71.43–94.38	50.46–92.06

alternative by focusing on specific parameters within the model. Only low-rank matrices are introduced in certain layers, primarily within the self-attention and feedforward blocks, allowing the majority of the pre-trained parameters to remain frozen. This technique reduces the memory and computational load while achieving task-specific adjustments by updating only the new, smaller matrices. By leveraging fine-tuning techniques for AV tasks and the efficiency gains of LoRA, exploring *V²LMs* as a solution to AV perception challenges has the potential to address critical limitations of existing methods, enhancing robustness against adversarial attacks while avoiding the performance degradation common in traditional defenses.

Solo vs. Tandem Design Comparison. Deploying *V²LMs* to enhance robustness against AEs in AVs raises important considerations for their implementation. Given the diverse range of tasks that AVs must perform, it is crucial to determine the best strategy for utilizing them. One approach involves using a separate *V²LM* for each specific AV task to improve robustness within each module, ensuring specialized and precise detection capabilities. This design, referred to as *Solo Mode*, is illustrated in 4a, where individual *V²LMs* are dedicated to tasks such as TSR, ALC, and VD. The formal representation of this design is:

$$Output_i = V^2LM^i_{solo}(Data_i, Prompt_i) \tag{4}$$

where $i \in \{TSR, ALC, VD\}$. Here, $Data_i$ denotes the dataset specific to each perception task T_i, and $Prompt_i$ is the prompt customized to fine-tune $V^2LM^i_{solo}$ for that particular task, ensuring task-specific optimization.

Alternatively, a single *V²LM* can handle all AV tasks using a tandem approach, providing a unified method to improve robustness across multiple modules, as depicted in 4b. In this design, multiple image query pairs, each corresponding to a different task, are combined into a single input: the images are concatenated using a separator and the queries are merged in the same order. The model processes this structured batch simultaneously within one forward pass, extracting task-specific outputs for each image-query pair independently while treating the collection as a cohesive input during execution:

$$\{Output_{TSR}, Output_{ALC}, Output_{VD}\} = V^2LM_{tandem}(Data_{all}, Prompt_{all}) \tag{5}$$

where $Data_{all}$ represents the combined dataset of all tasks, and $Prompt_{all}$ includes the concatenated prompts corresponding to each task. Moreover, a single *V²LM* could reduce memory and computational overhead, which is critical in the resource-constrained environment of AVs. In such systems, optimizing both efficiency and memory is vital, making a unified *V²LM* a more practical solution.

Integrating *V²LMs* with AV perception systems offers the potential to strengthen resilience against AEs in real-time. Currently, AV perception systems face significant limitations in mitigating attacks, which can lead to dangerous misinterpretations of sensor data. As Cao *et al.* [52] demonstrated, despite efforts to enhance the security of AV perception, significant vulnerabilities persist and traditional detection mechanisms often fail to mitigate these threats, leading to potentially dangerous consequences for AV decision-making and safety. By integrating *V²LM* in the end-to-end AV stack, AVs can benefit from its ability to process data, thereby working in parallel with perception tasks to enhance robustness and support accurate decision-making. The outputs of the perception system and *V²LM* can then be used by the downstream *planning* and *control* modules to act accordingly in the presence of AEs. This context is depicted in Figure 3.

It is crucial to evaluate the latency of *V²LM* integration to ensure it can enhance robustness against AEs in real-time without affecting the efficiency of the perception system. AV vision algorithms can process images at a rate of 2-4 frames per second (fps), which corresponds to approximately 250-500 ms per image [53].

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup

To evaluate the efficacy of *V²LMs*, the focus was on three critical tasks: Traffic Sign Recognition (TSR), Automated Lane Centering (ALC) and Vehicle Detection (VD). For the former, the German Traffic Sign Recognition Benchmark (GTSRB) was utilized, which includes images captured under various conditions such as different lighting and distances. Each picture from this dataset, which has 42 classes, was sent with the prompt *"Identify this traffic sign."* to the respective *V²LM*. For the ALC task, a dataset was generated using the CARLA simulator, which includes 6,000 training and 2,000 testing images. These images, taken from the

TABLE III

PERFORMANCE OF DNN MODELS ON BENIGN (B) AND UNSEEN ADVERSARIAL (A). DIFF IS THE ACCURACY/F1 DROP DUE TO ATTACKS.

Task	Model	Metric	A	B	Diff
TSR	YOLOv5-cls	Acc.	55.82%	94.88%	-39.06%
		F1-Score	58.10%	93.44%	-35.34%
	ViT-small	Acc.	23.12%	96.20%	-73.08%
		F1-Score	23.12%	96.36%	-73.24%
ALC	CLRerNet	Acc.	50.51%	90.91%	-40.40%
		F1-Score	44.86%	90.83%	-45.97%
VD	YOLOv5-dt	Acc.	62.27%	97.01%	-34.74%
		F1-Score	57.44%	96.60%	-39.16%

driver's perspective under various weather conditions and times of day, are classified into three categories indicating the next move: *Straight*, *Left*, and *Right*. The prompt used for this task was, *"As a car driver, at which direction should you turn the steering wheel?"*

For VD, CARLA was also used to create 7,000 training and 3,000 testing images. Captured from different viewpoints, under diverse weather conditions, at various times of day, and from different distances, these images are categorized into *Car presence* and *Car absence*. If a car is present, the bounding box coordinates are provided with x and y representing the center, and H and W indicating the height and width of the box. The prompt used for this task was, *"If a car is detected, provide the center coordinates and the dimensions of the bounding box for the car"*.

The evaluation began with assessing the VLMs' performance on zero-shot tasks across the test datasets. Subsequently, these VLMs were fine-tuned on AV-specific training data — yielding *V²LMs*— with their performance re-evaluated on the same test datasets to ensure comparability. The objective is to use a *V²LM* to enhance the ALC, TSR, and VD modules within the AV perception system against adversarial attacks. In the next step, AEs were generated from the same test datasets to assess robustness.

Although the fine-tuned models were trained on AV-specific data, the AEs remained ***unseen*** during training, allowing a reliable evaluation of each *V²LM*'s resilience to attacks on TSR, ALC, and VD. To achieve this, three types of black-box attacks were implemented. The first type of attack involves adversarial manipulation of traffic signs. In the study by Zhong *et al.* [24], shadows are utilized to conduct attacks on TSR algorithms, as shown in figure 2a.

This method employs shadows as a non-invasive mechanism to create physical AEs. By optimizing shadow properties such as shape and opacity using a differentiable renderer, the technique manipulates images under black-box conditions to induce misclassifications. It achieves a success rate of 90.47% on the GTSRB dataset, demonstrating its effectiveness and highlighting the vulnerabilities of current detection systems to such subtle manipulations.

The second type of attack targets the ALC mechanism of AVs. In the study by Sato *et al.* [6], the Dirty Road Patch (DRP) attack framework specifically targets ALC systems in AVs, exploiting vulnerabilities in deep learning-based lane detection. This method employs an optimization-based approach to systematically generate these patches, as illustrated in figure 2b, considering real-world conditions such as lighting and camera angles to ensure effectiveness across different environmental scenarios. The optimized DRPs cause the AV to make incorrect steering decisions, which were demonstrated to be highly successful in real-world driving scenarios with a success rate exceeding 97.5%

Figure 2c shows the third type of attack which focuses on the adversarial camouflage of vehicles. Zhou *et al.* [23] propose a physical adversarial attack known as the Robust and Accurate UV-map-based Camouflage Attack (RAUCA) to deceive VD algorithms such as YOLOv3 [54]. It employs a technique utilizing a differentiable neural renderer, which allows for the optimization of adversarial camouflages through gradient back-propagation, enhancing both the robustness and precision of the attacks under varying environmental conditions. Their method achieved an attack success rate of 97.48% on the target detection models, demonstrating the significant vulnerability of these systems to such sophisticated camouflage attacks.

B. RQ1: Fine-Tuning Increases Detection Performance

Table II shows VLMs' zero-shot performance on the test dataset, which performed poorly in ALC and TSR tasks, indicating difficulties in these specific AV applications. Although the models demonstrated decent performance in the VD task, this success may partly be due to their pre-training on large, diverse image datasets, which likely enhanced their general visual recognition capabilities. However, in the VD task, they struggled to detect vehicles in images taken at night, in the rain, or at long distances (highlighting their limitations under challenging environmental conditions despite strong general visual recognition).

To improve their performance, these models were fine-tuned using the same training datasets and prompts as before, and then their performance was re-evaluated with the same test dataset. Table IV shows notable accuracy improvements for all tasks after fine-tuning. For ALC, accuracy increased from 20.71%–50.91% to 86.61%–99.51%. For TSR, accuracy saw a substantial rise from 1.19%–31.24% to

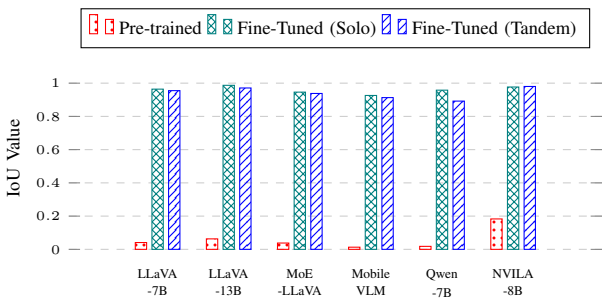


Fig. 5. IoU values before and after fine-tuning.

TABLE IV

COMPARISON OF V^2LM PERFORMANCE: BENIGN VS. AES. GREEN SHOWS IMPROVEMENT OF TANDEM OVER SOLO.

		Benign						AEs					
		Accuracy			F1-Score			Accuracy			F1-Score		
Task	**Model**	Solo	Tandem	Diff	Solo	Tandem	Diff	Solo	Tandem	Diff	Solo	Tandem	Diff
TSR	LLaVA-7B	95.93%	97.38%	1.45%	94.62%	91.22%	-3.40%	80.12%	86.51%	6.39%	86.44%	85.89%	-0.55%
	LLaVA-13B	97.15%	98.14%	0.99%	96.09%	93.21%	-2.88%	86.53%	89.01%	2.48%	86.08%	88.95%	2.87%
	MoE-LLaVA	95.24%	96.91%	1.67%	94.73%	95.78%	1.05%	80.03%	81.52%	1.49%	80.75%	82.34%	1.59%
	MobileVLM	88.19%	90.34%	2.15%	87.68%	90.18%	2.50%	79.57%	79.13%	-0.44%	77.55%	79.88%	2.33%
	Qwen-VL	95.80%	94.13%	-1.67%	95.78%	94.27%	-1.51%	82.83%	81.66%	-1.17%	83.16%	83.43%	0.27%
	NVILA-8B	99.04%	99.14%	0.1%	99.10%	99.20%	0.1%	86.04%	91.12%	5.08%	87.15%	91.10%	3.95%
ALC	LLaVA-7B	98.11%	95.41%	-2.70%	98.05%	95.24%	-2.81%	86.07%	84.83%	-1.24%	87.01%	85.38%	-1.63%
	LLaVA-13B	99.30%	98.31%	-0.99%	99.46%	97.85%	-1.61%	90.05%	86.69%	-3.36%	90.69%	87.38%	-3.31%
	MoE-LLaVA	96.86%	92.53%	-4.33%	96.23%	91.83%	-4.40%	82.54%	82.26%	-0.28%	83.27%	83.06%	-0.21%
	MobileVLM	84.41%	86.61%	2.20%	87.68%	89.53%	1.85%	78.21%	78.23%	0.02%	80.92%	81.21%	0.29%
	Qwen-VL	93.75%	93.91%	0.16%	93.83%	93.76%	-0.07%	88.81%	86.57%	-2.24%	89.48%	85.48%	-4%
	NVILA-8B	99.51%	99.41%	-0.1%	99.45%	99.52%	0.07%	99.25%	99.14%	-0.11%	99.13%	99.10%	-0.03%
VD	LLaVA-7B	95.84%	91.90%	-3.94%	94.56%	92.82%	-1.74%	89.23%	88.02%	-1.21%	87.84%	86.59%	-1.25%
	LLaVA-13B	97.05%	98.96%	1.91%	96.76%	95.97%	-0.79%	91.82%	90.09%	-1.73%	90.72%	90.41%	-0.31%
	MoE-LLaVA	95.52%	93.42%	-2.10%	94.78%	92.30%	-2.48%	88.46%	86.54%	-1.92%	87.16%	86.21%	-0.95%
	MobileVLM	86.61%	88.01%	1.40%	88.42%	89.38%	0.96%	80.13%	80.15%	0.02%	82.88%	82.97%	0.09%
	Qwen-VL	96.02%	95.01%	-1.01%	96.55%	96.17%	-0.38%	89.05%	89.06%	0.01%	90.15%	90.57%	0.42%
	NVILA-8B	99.93%	99.95%	0.02%	99.83%	99.84%	0.01%	99.81%	99.83%	0.02%	99.82%	99.83%	0.01%

90.34%–99.14%. In the VD task, accuracy improved from 60.64%–92.06% to 88.01%–99.95

These findings suggest that although they initially underperformed on AV tasks, fine-tuning them with relevant datasets can lead to substantial performance improvements, showing their potential utility in AV applications. Intersection over Union (IoU) measures the overlap between the predicted output and the ground truth in tasks like object detection. A higher IoU indicates more accurate localization, making it a crucial metric for evaluating model performance in AV perception tasks. After fine-tuning, the IoU values improved significantly, with `LLaVA-7B` increasing from 0.041 to 0.964, `LLaVA-13B-LoRA` from 0.063 to 0.987, `MoE-LLaVA` from 0.038 to 0.946, `MobileVLM` from 0.013 to 0.926, `Qwen-VL` from 0.018 to 0.9578, and `NVILA-8B` from 0.183 to 0.9768, as shown in Figure 5.

C. RQ2: V^2LMs Demonstrate Robustness under Attacks

Adversarial attacks pose a serious threat to deep learning based AV perception systems. Prior works have shown that DNN models suffer significant performance degradation across core AV tasks, including TSR, ALC, and VD, as summarized in Table V. For example, the GTSRB CNN model accuracy collapses to just 1.77% under the Shadow Attack [24], and even adversarial training, one of the most prominent defense methods, only modestly improves accuracy to 25.57%. Similarly, for ALC, OpenPilot ALC performance degrades dramatically to 2.50% under the DRP Attack [6], with established defense strategies such as JPEG compression [11], Gaussian noise addition [34], and autoencoder based denoising [55] achieving negligible improvements (around 3%). In VD, YOLOv3 experiences a drastic accuracy drop to 2.52% under RAUCA Attack [23], with no effective defense method proposed.

Building on these findings, we conduct our own robustness evaluation. To provide a fair and rigorous comparison, we evaluated traditional task-specific DNNs—`YOLOv5-cls` and `ViT-small` (TSR), `CLRerNet` (ALC), and `YOLOv5-dt` (VD)—on benign and previously *unseen* adversarial datasets (Table III). Our results confirm severe performance degradation, with accuracy reductions of 73.08% and 39.06% for TSR, 40.40% for ALC, and 32.91% for VD, clearly highlighting their vulnerability to novel attacks. In contrast, our evaluations of V^2LMs under the same unseen adversarial conditions reveal inherently superior robustness. Specifically, V^2LMs experienced substantially smaller accuracy drops: only 8.62%–15.81% for TSR, 4.94%–14.32% for ALC, and 5.23%–7.06% for VD, consistently maintaining high adversarial accuracy without additional defenses.

D. RQ3: Tandem V^2LMs Provide Similar Performance at Lower Memory Footprint

We evaluate whether one *tandem* V^2LM provides robustness across AV tasks comparable to separate *solo* V^2LMs. To assess this, VLMs are fine-tuned for three tasks, TSR, ALC, and VD, simultaneously using the same prompts. After fine-tuning, both solo and tandem V^2LMs were then evaluated on each task to determine its performance on the same test data. Table IV presents the evaluation results for the tandem design in non-adversarial scenarios, where a single V^2LM was trained to handle all three tasks simultaneously. The results reveal that the tandem design achieves high accuracy across perception tasks, often matching or surpassing the performance of solo models. This shows the effectiveness of the tandem design in maintaining robust performance across diverse tasks. Table IV further demonstrates the resilience of V^2LMs in AV tasks under adversarial conditions. `MobileVLM`, `MoE-LLaVA`, `LLaVA-7B`, `NVILA-8B`, `Qwen-VL`, and `LLaVA-13B-LoRA` allocated 6.03GB, 11.21GB, 13.56GB, 15.23GB, 16.58GB, and 26.15GB of storage, respectively. These results suggest that a single tandem V^2LM can generalize well across multiple tasks, providing robust performance comparable to the solo design, which requires 3x more storage for separate models. The tandem design offers a

TABLE V

DNNs ACCURACY UNDER ATTACKS AND DEFENSES.

Task	Model	Attack Type	Under-Attack Accuracy	Defense Strategy	Post-Defense Accuracy
TSR	GTSRB-CNN	Shadow [24]	1.77%	Adversarial Training [13]	25.57%
ALC	OpenPilot-ALC	DRP-Attack [6]	2.50%	JPEG Compression [11] Bit-Depth Reduction [33] Gaussian Noise [34] Median Blurring [33]	≈3% (no effective improvement)
VD	YOLOv3	RAUCA [23]	2.52%	No Defense Proposed	N/A

significant advantage in efficiency by keeping similar performance while requiring much less storage.

V. DISCUSSION

AV perception systems typically target latencies below 100 milliseconds to meet real-time operational requirements, especially in high-speed driving contexts [56]. In our experiments, the inference time t_{V^2LM} for `LLaVA-7B` was measured at 851 ms on an NVIDIA A100 GPU with 40 GB of VRAM, clearly exceeding the acceptable threshold t_{PS} for AV deployment. In contrast, the release of `NVILA` in late 2024 marked a significant improvement, reducing t_{V^2LM} to just 80 ms. This $10\times$ reduction highlights the rapid evolution of VLMs toward real-time readiness in AV perception pipelines. Despite this progress, deploying VLMs on embedded AV hardware remains challenging due to limitations in compute power and energy efficiency. High-performance models such as `LLaVA-7B` and `NVILA`, though effective on server-grade GPUs, often require substantial memory and parallel processing capabilities that are impractical for in-vehicle deployment. One potential solution to reduce hardware demands is quantization, a widely used model compression technique for LLMs that improves computational efficiency by converting high-precision data types to lower-precision formats [57].

This process significantly reduces memory usage and model size, making it more feasible to run VLMs on edge devices. However, it may also introduce quantization errors, which could degrade precision [58]. We applied quantization to `LLaVA-7B`, expecting lower resource usage. Although initially expected to reduce latency, the quantization approach did not yield the desired outcome; instead, when tested on the NVIDIA A100 40 GB, it exhibited latencies ranging from 0.851 s to 2.158 s and 11.657 s at full precision, 8-bit, and 4-bit levels, respectively. Based on our preliminary analysis, we found that the main source of this issue could be due to an increase in demand stemming from the additional operations required to maintain accuracy in image processing tasks.

VI. CONCLUSION

We present *V²LMs* as a novel approach to enhance the robustness of AV perception systems against adversarial attacks. We evaluate *Solo Mode* and *Tandem Mode*, and demonstrate that *V²LMs* maintain high adversarial accuracy without adversarial training while reducing storage requirements and maintaining comparable performance. Experimental results show that task-specific DNNs suffer performance drops of 33%–74% under adversarial attacks, whereas *V²LMs* have reductions of less than 8% on average.

ACKNOWLEDGMENT

We gratefully acknowledge the support provided by the U.S. Department of Transportation (DOT) through the National Center for Transportation Cybersecurity and Resiliency (TraCR) under Grant No. 69A3552344812-2027534 and 69A3552348317. This work has also been partially supported by NSF under grant CNS-2443252 and The BMW Group. The authors also appreciate the support from Google GCP Credit Award program.

REFERENCES

[1] L.-H. Wen and K.-H. Jo, "Deep learning-based perception systems for autonomous driving: A comprehensive survey," *Neurocomputing*, vol. 489, pp. 255–270, 2022.

[2] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 187–210, 2018.

[3] S. M. Peyghambarzadeh, F. Azizmalayeri, H. Khotanlou, and A. Salarpour, "Point-planenet: Plane kernel based convolutional neural network for point clouds analysis," *Digital Signal Processing*, vol. 98, p. 102633, 2020.

[4] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 159–15 168.

[5] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu, "Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems," *arXiv preprint arXiv:2201.06192*, 2022.

[6] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 3309–3326.

[7] M. Soltani, M. Davari, M. Bahadori, A. Kokhahi, M. Bahadori, and M. Soleimani, "Explainable artificial intelligence-based machine analytics and deep learning in medical science," in *Explainable Artificial Intelligence in Medical Imaging*. Auerbach Publications, 2025, pp. 205–219.

[8] M. Saberian, N. Zafarmomen, A. Neupane, K. Panthi, and V. Samadi, "Hydroquantum: A new quantum-driven python package for hydrological simulation," *Environmental Modelling & Software*, p. 106736, 2025.

[9] P. Afshin, D. Helminiak, T. Lu, T. Yen, J. M. Jorns, M. Patton, B. Yu, and D. H. Ye, "Breast cancer classification in deep ultraviolet fluorescence images using a patch-level vision transformer framework," *arXiv preprint arXiv:2505.07654*, 2025.

[10] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[11] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *arXiv preprint arXiv:1711.00117*, 2017.

[12] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International conference on machine learning*. PMLR, 2018, pp. 5286–5295.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[14] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arxiv 2023," *arXiv preprint arXiv:2308.12966*, vol. 1, no. 8, 2023.

[15] Z. Liu, L. Zhu, B. Shi, Z. Zhang, Y. Lou, S. Yang, H. Xi, S. Cao, Y. Gu, D. Li *et al.*, "Nvila: Efficient frontier visual language models," *arXiv preprint arXiv:2412.04468*, 2024.

[16] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, "Adversarial training can hurt generalization," *arXiv preprint arXiv:1906.06032*, 2019.

[17] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec *et al.*, "YOLOv5: PyTorch implementation of YOLO object detector," https://github.com/ultralytics/yolov5, 2020.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *arXiv preprint arXiv:1805.12152*, 2018.

[20] X. Chu, L. Qiao, X. Lin, S. Xu, Y. Yang, Y. Hu, F. Wei, X. Zhang, B. Zhang, X. Wei *et al.*, "Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices," *arXiv preprint arXiv:2312.16886*, 2023.

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[22] B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Zhang, M. Ning, and L. Yuan, "Moe-llava: Mixture of experts for large vision-language models," *arXiv preprint arXiv:2401.15947*, 2024.

[23] J. Zhou, L. Lyu, D. He, and Y. Li, "Rauca: A novel physical adversarial attack on vehicle detectors via robust and accurate camouflage generation," *arXiv preprint arXiv:2402.15853*, 2024.

[24] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 345–15 354.

[25] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles," *IEEE Intelligent Transportation Systems Magazine*, 2024.

[26] M. Aldeen, P. MohajerAnsari, J. Ma, M. Chowdhury, L. Cheng, and M. D. Pesé, "An initial exploration of employing large multimodal models in defending against autonomous vehicles attacks," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 3334–3341.

[27] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

[28] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[29] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li *et al.*, "Sok: On the semantic ai security in autonomous driving," *arXiv preprint arXiv:2203.05314*, 2022.

[30] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 176–194.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[33] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[34] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," *arXiv preprint arXiv:1711.01991*, 2017.

[35] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics gems IV*, 1994, pp. 474–485.

[36] N. Triki, M. Karray, and M. Ksantini, "A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles," *Applied Sciences*, vol. 13, no. 8, p. 4793, 2023.

[37] Y. Dubey, Y. Tarte, N. Talatule, B. Sable, C. Taywade, and R. Umate, "An artificial intelligence based autonomous road lane detection and navigation system for vehicles," 2024.

[38] C. Caraffi, T. Vojíř, J. Trefnỳ, J. Šochman, and J. Matas, "A system for real-time detection and tracking of vehicles from a single car-mounted camera," in *2012 15th international IEEE conference on intelligent transportation systems*. IEEE, 2012, pp. 975–982.

[39] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[40] Z. Lu, "A theory of multimodal learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[41] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," *arXiv preprint arXiv:2310.03986*, 2023.

[42] Z. Zhao, E. Monti, J. Lehmann, and H. Assem, "Enhancing contextual understanding in large language models through contrastive decoding," *arXiv preprint arXiv:2405.02750*, 2024.

[43] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[45] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[46] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[47] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2818–2829.

[48] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[49] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[50] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[51] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[52] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 2267–2281.

[53] K. Blachut, M. Danilowicz, H. Szolc, M. Wasala, T. Kryjak, and M. Komorkiewicz, "Automotive perception system evaluation with reference data from a uav's camera using aruco markers and dcnn," *Journal of Signal Processing Systems*, vol. 94, no. 7, pp. 675–692, 2022.

[54] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[55] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.

[56] T. Jin, W. Ding, M. Yang, H. Zhu, and P. Dai, "Benchmarking perception to streaming inputs in vision-centric autonomous driving," *Mathematics*, vol. 11, no. 24, p. 4976, 2023.

[57] IBM. (2023) What is quantization? Accessed: 2023-11-14. [Online]. Available: https://www.ibm.com/think/topics/quantization

[58] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, "Awq: Activation-aware weight quantization for on-device llm compression and acceleration," *Proceedings of Machine Learning and Systems*, vol. 6, pp. 87–100, 2024.