Camouflage Attack on Vision-Language Models for Autonomous Driving

Dehong Kong[1,2], Sifan Yu[1,2],Linchao Zhang[3], Shirui Luo[3], Siying Zhu[3],
Yanzhao Su[4], WenQi Ren[1,2,5,†]

[1] School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University

[2]MoE Key Laboratory of Information Technology

[3]Information Science Academy,China Electronics Technology group Coporation

[4] Rocket Force University of Engineering [5]Guangdong Provincial Key Laboratory of Information Security Technology

Abstract

Vision-Language Models for Autonomous Driving (VLM-AD) is gradually becoming a research hotspot. VLM enhances the performance of AD systems with excellent reasoning capabilities. However, it faces critical safety risks to adversarial attacks. While current research predominantly focuses on digital adversarial attacks, physical attacks against VLM-AD remain underexplored. In this paper, we introduce the first Camouflage Attack framework CAM-VLAD on Vision-Language Models for Autonomous Driving. Leveraging the vulnerability of some specific layers, we design a novel method Feature Divergence Attack. Modeling physical camouflage textures as differentiable parameters, we apply the adversarial attack in the vision-language embedding, bypassing traditional pixel-level optimization. The attack generates physically realizable textures to maximize feature distortion and mislead the decision of VLM-AD. Experiments show our method achieves strong attack performance making VLM-AD generate many wrong driving instructions.

1. Introduction

Vision-Language Models (VLMs) are becoming pivotal decision-making systems for autonomous vehicles, combining visual perception with language understanding to enable multimodal reasoning. These systems allow vehicles to interpret driving instructions, generate explainable decisions, and interact naturally with humans through dialogue. However, this multimodal fusion introduces significant security risks. Studies reveal heightened vulnerabilities of VLM-AD to adversarial attacks, such as reversing directional commands. Recent advances in adversarial attacks on VLM-AD are significant. CAD [5] targets low-level reasoning breakdown by generating and injecting deceptive semantics. ADvLM [7] introduces Scenario-Associated Enhancement, an approach where attention mechanisms select key frames and perspectives within driving scenarios to optimize adversarial perturbations. However, these approaches cannot be effectively deployed in real-world physical environments.

One of the predominant physical attack methods is patch-based attack but it remains limited in robustness due to the multi-sensor configuration and diverse viewing angles inherent in autonomous driving systems. Furthermore, attack loss for object detectors cannot be effectively applied in VLM-AD.

We propose a new camouflage attack framework incorporating feature divergence loss to disturb the feature space of VLM-AD. Unlike traditional camouflage attacks, our method is designed for the architecture of VLM-AD. Specifically, CAM-VLAD targets the inherent vulnerabilities in multimodal feature alignment, generating physically realizable adversarial camouflage textures. This approach exploits the feature vulnerabilities in autonomous driving systems to generate adversarial camouflage texture. Specifically, we consider attacking the visual feature of the encoder and projector layers to deviate from the original feature space. We also observe that the distribution of multiple features in the output of the final hidden layer is sensitive to changes in the camouflage textures, which inspires us to introduce a variance-guided penalty to restrict the direction of optimization. Our framework achieves an end-to-end feature-based attack pipeline, generating physical-world adversarial textures that simultaneously deceive multimodal semantic reasoning.

In summary, the main contributions list as follows:(1)We are the first to introduce camouflage attacks on VLM-AD. (2)By analyzing the shortcomings of existing approaches, an innovative feature-based attack is proposed to target VLM-AD. (3)Experimental results demonstrate that our proposed attack method achieves strong performance and excellent transferability.

† Corresponding Author.

2. Related Work

VLMs in Autonomous Driving. Recent advances in large language models (VLMs) have expanded their applications in autonomous driving. CODA-LM [1] introduced an automated benchmark for long-tail driving scenarios, using text-based VLMs for evaluation and demonstrating enhanced decision analysis via structured prompts. Dolphins [3] developed a chain-of-thought reasoning framework for multimodal interaction, enabling real-time learning and self-correction through driving-instruction fine-tuning. OmniDrive [6] proposed a sparse query-based architecture for 3D scene modeling, integrating dynamic-static object representation with memory-enhanced positional encoding.

Physical Adversarial Attacks. Physical adversarial attacks manipulate object characteristics to deceive vision systems, categorized into patch-based and camouflage-based approaches. Patch-based methods apply localized adversarial patterns to object surfaces or backgrounds. Those methods are mainly designed to attack object detectors. DGA [8] propose a new direction-guided attack to deceive real-world aerial detectors. However, their planar constraints limit robustness under multi-view or long-distance conditions. Camouflage-based methods enhance stealth by optimizing 3D textures or shapes. FCA [4] introduced Full-coverage Camouflage Attack, which maps adversarial textures onto entire vehicle surfaces using neural rendering and environmental transformations to address multi-view failures.

Adversarial Attacks on VLM-AD. Adversarial attacks on VLMs for autonomous driving systems have attracted significant attention, focusing on dynamic scene adaptability, multimodal vulnerabilities, and robustness in safety-critical scenarios. Zhang et al. [7] developed ADvLM, employing semantic-invariant induction to create instruction libraries and scene-correlation optimization for temporal perturbations, enhancing attack robustness across dynamic perspectives. For black-box scenarios, Wang et al. [5] proposed the Cascaded Adversarial Disturbance (CAD) framework, inducing cross-reasoning-chain errors via decision-chain disruption and risk-scenario induction in dynamic environments.

3. Method

Figure 1 shows the framework to attack VLM-AD. The attack scheme is to generate the adversarial camouflage texture utilizing the neural renderer to paint on the surface of the 3D vehicle model. Based on the analysis of the vulnerability of VLM, we manipulate the intermediate features of vision models, projectors, and hidden layers of VLM to disturb the output of models.

3.1. Problem Formulation

Given a training dataset ($\mathbf{X}$,θ_c) where $\mathbf{X}$ and θ_c are the sampled images and the corresponding camera parameters respectively, a 3D car model with a mesh $\mathbf{M}$ and a texture $\mathbf{T}$, 2D car image $\mathbf{O}$ can be generated by a renderer $\mathcal{R}$:

$$\mathbf{X}_{\mathrm{T}} = \mathcal{R}(\mathbf{M}, \mathbf{T}; \theta_c). \tag{1}$$

To realize the adversarial camouflage attack, we replace the origin texture $\mathbf{T}$ with adversarial texture $\mathbf{T}_{adv}$ and obtain the adversarial image $\mathbf{X}_{\mathrm{T}_{adv}}$ with transformation function ϕ. We aim to input ($\mathbf{X}_{adv}$,t) to attack $\mathcal{F}$ to output the wrong text or reduce its performance where $\mathcal{F}$ is VLM-AD and t is the benign text input.

We treat the manipulation as an optimization problem, and the function is expressed as follows

$$\hat{\mathbf{T}}_{adv} = \underset{\mathrm{T_{adv}}}{\arg\max}\, \mathcal{J}(\mathcal{F}(\phi(\mathbf{X}_{\mathrm{T}}), t), \mathcal{F}(\phi(\mathbf{X}_{\mathrm{T}_{adv}}), t)), \tag{2}$$

where $\hat{\mathbf{T}}_{adv}$ is the trained adversarial texture and $\mathcal{J}(\cdot,\cdot)$ is the loss function.

3.2. Camouflage Attack

We generate the adversarial camouflage texture by utilizing a differentiable neural renderer. It enables the direct application of customized textures onto 3D car models. This is the first attempt in the field of autonomous driving adversarial attacks. To ensure the naturalness of the generated adversarial camouflage, we utilize the smooth loss to reduce the inconsistency among adjacent pixels. For a rendered car image painted with adversarial camouflage $\mathbf{X}_{adv}$, the calculation of smooth loss can be written as

$$\mathcal{L}_{\text{smooth}} = \sum_{i,j} \left((x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2 \right), \tag{3}$$

where $x_{i,j}$ is the pixel value of $\mathbf{X}_{adv}$ at coordinate (i, j).

3.3. Feature-Divergence

The VLM-AD mainly consists of an encoder, a projection layer, and internal hidden layers. Inspired by VLM-based digital attacks[2], we consider a multi-layer feature attack that causes the features before and after the layers to deviate.

The outputs of the encoder and projection layers are more worthy of attention compared to the intermediate layers of the VLM, for the following two reasons: 1) The intermediate layers of the VLM are greatly influenced by text

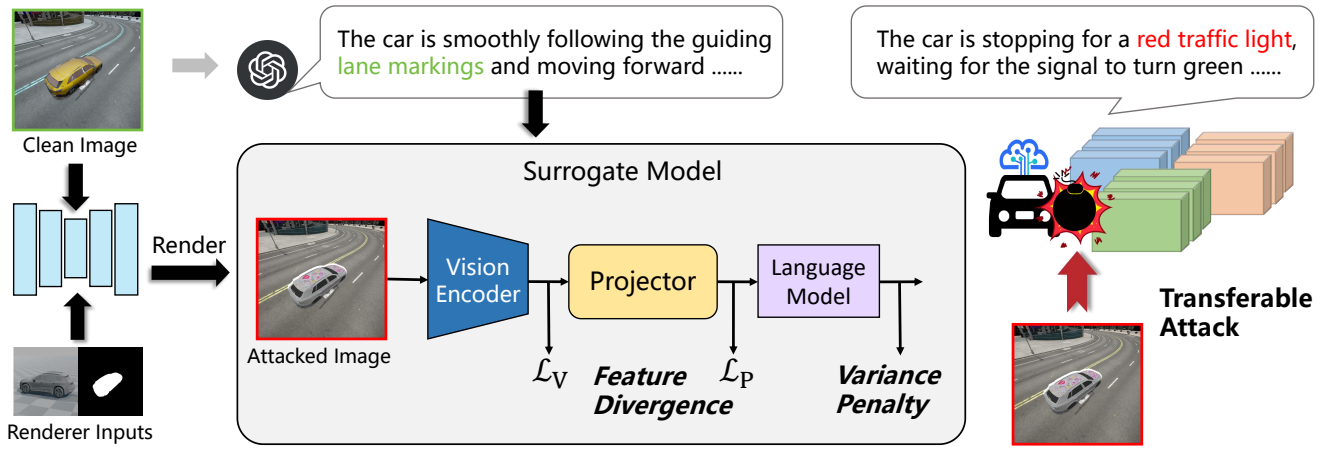

Figure 1. Attack Framework CAM-VLAD. Our approach introduces Feature-Divergence which targets to attack feature space of the encoder and projector. Variance-Penalty is proposed to guide attack direction towards variance increment.

input and visual information is conditional guidance. 2) Attacks on the encoder and projection layers exhibit better transferability. we extract spatiotemporal representations of $\mathbf{X}$ and $\mathbf{X}_{\mathrm{T_{adv}}}$ using the vision encoder $\mathcal{F}_V$. To make the features diverse, we optimizes the adversarial texture by minimizing the cosine similarity between the original and the adversarial features:

$$\mathcal{L}_V = \sum_{i=1}^{N_1} \frac{\cos(\mathcal{F}_V(X)_i, \mathcal{F}_V(X_{\mathrm{T_{adv}}})_i)}{N_1}, \tag{4}$$

where $\mathcal{F}_V(\mathbf{X})_i$ and $\mathcal{F}_V(\mathbf{X}_{\mathrm{T_{adv}}})_i)$ is the i-th feature of $\mathbf{X}$ and $\mathbf{X}_{\mathrm{T_{adv}}}$, $\cos(\cdot,\cdot)$ is cosine similarity loss. Similarly, the projector attack can be implemented by minimizing the following projector feature loss:

$$\mathcal{L}_P = \sum_{i=1}^{N_2} \frac{\cos(\mathcal{F}_{VP}(X)_i, \mathcal{F}_{VP}(X_{\mathrm{T_{adv}}})_i)}{N_2}, \tag{5}$$

where $\mathcal{F}_{VP}(\cdot)$ is vision encoder and projector function.

Variance-Penalty. The above attack is in the direction of deviating from the feature space of the original sample. However, without constraints, the attack may become unstable. Therefore, we consider constructing a penalty term on the output of the last hidden layer, encouraging the feature distribution of the image to deviate towards increased variance, thus making the model easier to mislead. Penalty term can be calculated as follows:

$$\mathcal{P} = \min_{1 \le i \le n} \mathrm{Var}(\mathcal{F}_h^i(\mathbf{X}_{\mathrm{T}_{adv}})), \tag{6}$$

where Var is the variance of features, $\mathcal{F}_h(\mathbf{X}_{\mathrm{T}_{adv}})$ is the output of the last hidden layer, n is the number of features.

Total loss can be written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{smooth}} + \alpha\mathcal{L}_{\mathrm{V}} + \beta\mathcal{L}_{\mathrm{P}} - \lambda\mathcal{P}, \tag{7}$$

where α, β and λ are the weights to balance the contribution of loss term.

4. Experiment

4.1. Experiment Setup

We select SoTA VLM-based AD models Dolphins for attack. We implement patch attack DGA and a camouflage attack FCA [4] for AD to be baseline methods. We follow the common practice metrics in relevant works for comparison: CODA-LM [1] uses text-only VLMs, *e.g.* GPT-4, as evaluators to score model responses. OmniDrive [6] employs rule-based language metrics to evaluate sentence similarity at the word level. We follow FCA to utilize photo-realistic datasets to perform the experiments. We select the simulator CARLA for AD research. The CARLA simulator provides a variety of high-fidelity digital scenarios. We use the same datasets with FCA directly, the training set consists of 12,500 high-resolution images. Learning rate and max epoch is 0.1 and 5 respectively.

4.2. Attacking performance

In this experiment, we evaluate the effectiveness of our method against VLM-AD, with a focus on performance metrics from both the VLM-Judge and NLP metrics. The results are compared with various existing attack methods. As shown in Table 1, our method performs strong performance and We lead in all metrics and significantly outperform other physical attack methods. Notably, in the average of NLP metric, our method achieves 48.30%, which is comparable to FCA's 54.64%, leading by 6.34%.

Type	Metrics	Noise	DGA[8]	FCA[4]	Ours
VLM-Judge	General↓	10.00	8.20	7.80	**6.20**
	Regional↓	10.00	8.60	7.60	**6.40**
	Suggestion↓	10.00	7.10	7.90	**6.30**
	Average↓	10.00	7.97	7.77	**6.30**
NLP metrics	BLEU↓	100%	57.77%	53.99%	**48.45%**
	METEOR↓	100%	66.01%	49.22%	**42.16%**
	ROUGE↓	100%	78.26%	60.71%	**54.30%**
	Average↓	100%	67.34%	54.64%	**48.30%**

Table 1. Attack performance comparison with patch-base attack and camouflage attack

	BLEU	METEOR	ROUGE	Average
$\mathcal{L}_V$	73.90%	70.17%	77.70%	73.92%
$\mathcal{L}_P$	52.02%	46.25%	58.45%	52.24%
$\mathcal{L}_P + \mathcal{L}_V$	49.26%	43.36%	55.55%	49.39%
$\alpha\mathcal{L}_P + \beta\mathcal{L}_V - \lambda\mathcal{P}$	**48.45%**	**42.16%**	**54.30%**	**48.30%**

Table 2. Ablation study on different loss.

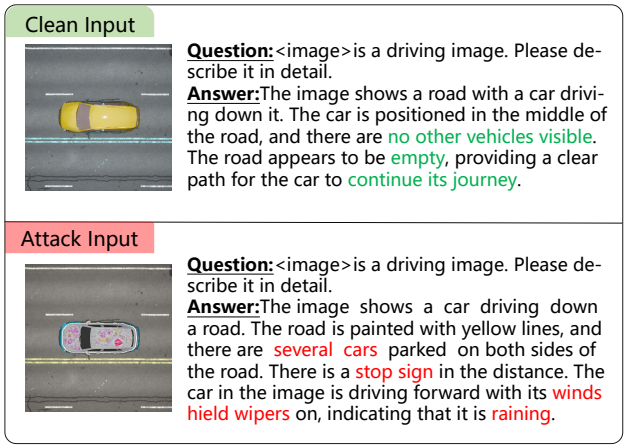

Figure 2. Visualization Before and After Attack.

Figure 2 shows an attack case. In the pre-attack output, the VLM-AD accurately describes the image of a road with a car driving on it. The green highlights indicates the original keywords, all of which are relevant and correct based on the image content. After the camouflage attack, the description of VLM-AD significantly deviates from the original content. Key errors are highlighted in red, such as the statement that "several cars are parked on both sides of the road" and the image "indicating that it is raining." These errors are clearly incorrect based on the original image, where there were no cars parked and the weather conditions are not depicted as rainy. Lastly, it can be seen that VLM-AD tends to output longer texts, which indicates that our attack has affected the performance.

Ablation Study. The results in Table 2 show the impact of different loss functions on the attack performance. Encoder and projector attack shows great attacking performance, leading far beyond the single attack. With the variance-guided penalty, our method achieves the best scores, demonstrating the most effective attack by successfully manipulating the outputs. These results confirm that our designed attack strategy significantly impacts the model's performance.

5. Conclusion

We propose a novel camouflage attack CAM-VLAD to disturb features of the VLM-base autonomous driving system. The experiment demonstrates the effectiveness of our attack method and its applicability to physical environments.

References

[1] Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 2, 3

[2] Linhao Huang, Xue Jiang, Zhiqiang Wang, Wentao Mo, Xi Xiao, Bo Han, Yongjie Yin, and Feng Zheng. Image-based multimodal models as intruders: Transferable multimodal attacks on video-based mllms. *arXiv preprint arXiv:2501.01042*, 2025. 2

[3] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *ECCV*, 2024. 2

[4] Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Zhiqiang Gong, Xiaoya Zhang, Wen Yao, and Xiaoqian Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *AAAI*, 2022. 2, 3, 4

[5] Lu Wang, Tianyuan Zhang, Yang Qu, Siyuan Liang, Yuwei Chen, Aishan Liu, Xianglong Liu, and Dacheng Tao. Black-box adversarial attack on vision language models for autonomous driving. *arXiv preprint arXiv:2501.13563*, 2025. 1, 2

[6] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024. 2, 3

[7] Tianyuan Zhang, Lu Wang, Xinwei Zhang, Yitong Zhang, Boyi Jia, Siyuan Liang, Shengshan Hu, Qiang Fu, Aishan Liu, and Xianglong Liu. Visual adversarial attack on vision-language models for autonomous driving. *arXiv preprint arXiv:2411.18275*, 2024. 1, 2

[8] Yue Zhou, Shuqi Sun, Xue Jiang, Guozheng Xu, Fengyuan Hu, Ze Zhang, and Xingzhao Liu. Dga: Direction-guided attack against optical aerial detection in camera shooting direction-agnostic scenarios. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–22, 2024. 2, 4