Visual Adversarial Attack on Vision-Language Models for Autonomous Driving

Tianyuan Zhang[1], Lu Wang[1], Xinwei Zhang[1], Yitong Zhang[1], Boyi Jia[1],
Siyuan Liang[2], Shengshan Hu[3], Qiang Fu[1], Aishan Liu[1], Xianglong Liu[1]
[1]Beihang University, China
[2]National University of Singapore, Singapore
[3]Huazhong University of Science and Technology, China

Abstract

Vision-language models (VLMs) have significantly advanced autonomous driving (AD) by enhancing reasoning capabilities. However, these models remain highly vulnerable to adversarial attacks. While existing research has primarily focused on general VLM attacks, the development of attacks tailored to the safety-critical AD context has been largely overlooked. In this paper, we take the first step toward designing adversarial attacks specifically targeting VLMs in AD, exposing the substantial risks these attacks pose within this critical domain. We identify two unique challenges for effective adversarial attacks on AD VLMs: the variability of textual instructions and the time-series nature of visual scenarios. To this end, we propose ADvLM, the first visual adversarial attack framework specifically designed for VLMs in AD. Our framework introduces Semantic-Invariant Induction, which uses a large language model to create a diverse prompt library of textual instructions with consistent semantic content, guided by semantic entropy. Building on this, we introduce Scenario-Associated Enhancement, an approach where attention mechanisms select key frames and perspectives within driving scenarios to optimize adversarial perturbations that generalize across the entire scenario. Extensive experiments on several AD VLMs over multiple benchmarks show that ADvLM achieves state-of-the-art attack effectiveness. Moreover, real-world attack studies further validate its applicability and potential in practice.

1. Introduction

Owing to their strong generalization capabilities and inherent interpretability, vision-language models (VLMs) have demonstrated exceptional performance across various tasks, including autonomous driving (AD). By enabling autonomous systems to comprehend scenarios and process natural language, VLMs could serve as the brain and offer effective solutions for advanced reasoning in complex scenarios and more efficient human-machine interaction [3, 40, 45, 55]. As a novel solution in end-to-end AD, VLMs present significant potential for future development.

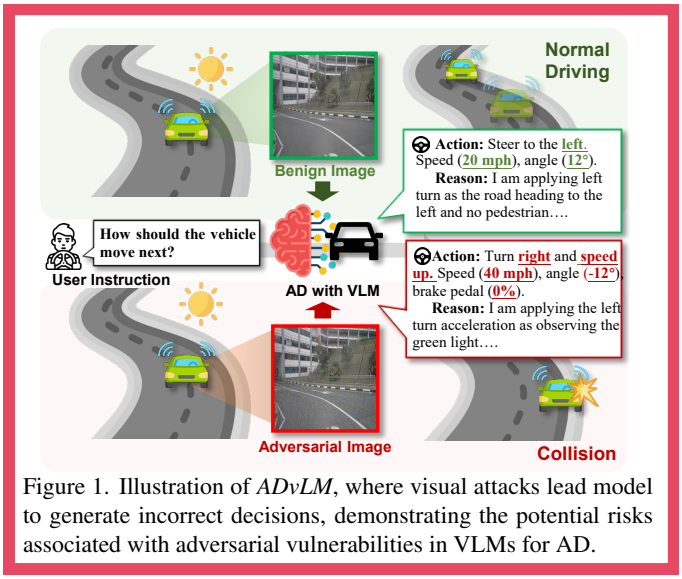


Figure 1. Illustration of *ADvLM*, where visual attacks lead model to generate incorrect decisions, demonstrating the potential risks associated with adversarial vulnerabilities in VLMs for AD.

However, VLMs exhibit significant vulnerabilities and lack robustness, particularly when faced with carefully crafted visual perturbations such as adversarial attacks [11, 14, 15, 17–20, 26, 32, 34–37, 52, 53, 56–58]. While various attack methods have been proposed, most existing research focuses on general VLMs and has not specifically addressed the unique requirements of AD. Identifying and addressing these vulnerabilities is essential in safety-critical domains like AD, as failures in VLMs could lead to severe consequences, including accidents or compromised decision-making.

In this paper, we take the first step to study adversarial attacks on VLM for AD. However, it is highly non-trivial to simply extend current adversarial attacks on general VLMs to this scenario, where We posit two key challenges unique

in VLM for AD as follows. ❶ Attack should work among varied textual instructions with different phrases/sentences that convey the same task semantics. ❷ Attack should work for a specific time-series driving scenario with multiple visual frames and perspective shifts. To address these challenges, we propose *ADvLM*, the first visual adversarial attack framework specifically tailored for VLMs in autonomous driving. In the textual modality, we propose the Semantic-Invariant Induction where we construct a low-semantic-entropy prompts library containing diverse textual instructions with the same semantics. Specifically, we employ a large language model to generate prompt variants from a seed and then refine them to promote the diversity in expressions guided by semantic entropy. In the visual modality, we introduce Scenario-Associated Enhancement, where we select critical frames/perspectives within the driving scenario based on model attentions, and further optimize the adversarial perturbations based on the pivotal frames while traversing the prompts library, such that the attack can generalize over the whole scenario. In this way, we can generate adversarial attacks across an expanded text and image input space, resulting in attacks that can remain effective and induce targeted behaviors across both varied instructions and time-series viewpoints in VLMs for AD.

To demonstrate its efficacy, we conduct extensive experiments on several VLMs for AD over multiple datasets, where our attack significantly outperforms other baselines with the highest Final Score reduction (+16.97% and 7.49%) in both white-box and black-box settings. In the closed-loop evaluation associated with the simulation environment CARLA, *ADvLM* also proves most effective, yielding a Vehicle Collisions Score of 2.954. In addition, we conduct real-world studies on physical vehicles to further demonstrate the potential of our attacks. Our **contributions** are shown as:

- We propose *ADvLM*, the first adversarial attack specifically designed for VLMs in AD, addressing the unique challenges inherent in AD.
- We introduce Semantic-Invariant Induction in the textual domain and Scenario-Associated Enhancement in the visual domain, ensuring attack effectiveness across varied instructions and sequential viewpoints.
- Extensive experiments in both the digital and physical worlds demonstrate that *ADvLM* outperforms existing methods and shows high potential in practice.

2. Related Works

Adversarial Attacks on VLMs. With the widespread deployment and outstanding performance of VLMs in multimodal question answering and reasoning, their robustness [21, 29, 58] has gradually attracted attention in recent years. Prior to our work, researchers have explored adversarial attacks against general VLMs. Due to the multimodal nature of VLMs, most adversarial attacks involve perturbations applied simultaneously to both image and text modalities. Drawing inspiration from adversarial attacks in vision tasks [10, 16, 22–28, 33, 48, 59, 64, 65], these methods [12, 13, 47, 49, 54, 60, 62] typically rely on end-to-end differentiable gradients. [60] introduced the first multimodal adversarial attack on VLMs, which paved the way for subsequent attacks that began exploring more practical black-box settings [7, 56, 67]. Researchers typically aim for attack methods that introduce minimal perturbations while having a strong impact, leading some studies to focus solely on attacking the visual modality of VLMs. [2, 61, 63, 66] demonstrate that it is possible to attack specific targets using only image-based perturbations successfully. Adversarial attacks that target only the text modality are uncommon in VLMs, as they primarily focus on large language models.

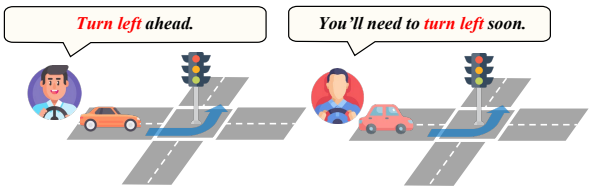
Despite the development of various adversarial attack techniques for general VLMs, there is a notable lack of methods specifically addressing the robustness of VLMs in the safety-critical context of AD.

VLMs in Autonomous Driving. Recent research has increasingly focused on VLMs as a means to tackle AD tasks by integrating both visual and linguistic inputs. These models excel in tasks like perception, reasoning, and planning, which are essential for AD systems. The tasks of AD VLMs can be primarily categorized into two types. The first is the core function of VLMs, namely VQA, such as the classic Reason2Drive [42], LingoQA [41] and Dolphins [38]. These foundational works thoroughly explore the enhanced role of VLMs in AD, particularly their meticulous reasoning and explanatory abilities in various driving-related tasks such as scene understanding, behavior prediction, and dialogue. The second is driving planning or control, closely related to the operations of AD. GPT-Driver [40], Driving with LLMs [3], and MTD-GPT [31] pioneered improvements to VLMs for driving planning. However, these works only considered the driving problem in open-loop settings, overlooking issues such as cumulative errors and end-to-end interpretability. In contrast, LMDrive [45] is the first to propose a VLM-based driving method within closed-loop settings, addressing these critical limitations. Other methods have integrated VQA and planning/control within VLM frameworks, offering a more holistic approach to AD. DriveLM [46], DriveMLM [50],m and DriveGPT4 [55] all go beyond basic conversations to implement more refined driving control and decision-making reasoning.

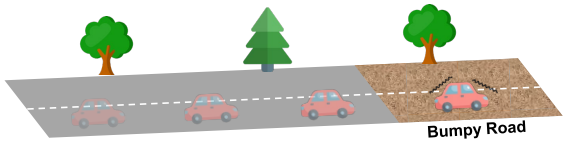
This paper selects representative models from each of the three categories for a comprehensive robustness analysis.

3. Problem and Motivation

Attack on VLMs. Adversarial attacks on VLMs for AD aim to manipulate a model's output by introducing carefully crafted perturbations into the input data. Specifically, an ad-



(a) Variability of textual instructions



(b) Time-series nature of visual scenarios

Figure 2. Illustration of the main challenges of attacks for VLMs in AD including the variability of textual instructions and the time-series nature of visual scenarios.

versary applies adversarial perturbations to a benign query $(\mathbb{V}, t)$, where $\mathbb{V} \subset \mathbb{V}_{\text{all}}$ denotes a specific sequence of visual inputs in AD, representing multiple frames rather than a single image. Here, $\mathbb{V}_{\text{all}}$ is the domain of all possible visual inputs for the model. This results in an adversarial query $\mathcal{P}(\mathbb{V}, t)$, where $\mathcal{P}(\cdot)$ denotes the adversarial perturbation function. The goal of $\mathcal{P}$ is to induce the VLM, denoted F_θ, to output a targeted or undesirable response y^* instead of the intended benign response. This manipulation is formally defined by maximizing the likelihood of the response y^* under the adversarial input:

$$\max_{\mathcal{P}} \log p(y^* | \mathcal{P}(\mathbb{V}, t)), \tag{1}$$

where p represents the probability function $F_\theta : \mathbb{Q} \to \mathbb{R}$, with $\mathbb{Q} = \mathbb{V}_{\text{all}} \times \mathbb{T}$ as the input query domain, comprising sequences of visual inputs $\mathbb{V} \subset \mathbb{V}_{\text{all}}$ and textual inputs $t \in \mathbb{T}$, and $\mathbb{R}$ as the response domain. In this work, we primarily focus on *attacks in the visual domain*, ensuring generated perturbations remain consistent within the same sequence.

Challenges and Attacking Goals. Common VLM adversarial attacks focus on fixed inputs (*i.e.*, specific textual and visual input), but AD introduces unique challenges that require tailored approaches for effective attacks. We identify two key challenges essential for effective adversarial attacks in AD (as shown in Fig. 2), which differ this attack from those for general VLMs.

❶ Variability of textual instructions. Drivers in AD often use varied textual instructions with different phrases for the same task, such as "turn left at the intersection" and "turn left ahead". In other words, these instructions are shown in different phrases but convey the same semantics and intent. To ensure a stable attack, visual perturbations δ must remain effective across an expanded set of semantically equivalent prompts $\widetilde{\mathbb{T}}$, derived from an original prompt, leading the VLM to consistently produce an incorrect response y^* across all prompts in $\widetilde{\mathbb{T}}$ that convey the same command.

❷ Time-series nature of visual scenarios. When driving, the vehicle's perspective shifts frequently due to movement and environmental factors. AD models must adapt to visual changes from motion and temporal dependencies. Unlike static tasks, given an instruction, attacking AD VLMs demands the perturbations to make reliable impacts on a series of frames even as perspectives and image quality vary. Let $\widetilde{\mathbb{V}}$ represent a collection of different perspectives generated from an original frame in $\mathbb{V}$, capturing the series of frames typical of time-series visual scenarios. This formulation ensures that the adversarial attack is effective across a dynamic visual sequence in an AD context.

To sum up, the adversary should consider generating adversarial perturbations that induce the AD VLM to produce the targeted response y^* consistently as follows:

$$\delta = \arg\max \sum_{\mathbb{V}} \sum_{\widetilde{\mathbb{T}}} \log p(y^* | \aleph(\widetilde{\mathbb{V}}, \delta), t), \tag{2}$$

where $t \in \widetilde{\mathbb{T}}$ represents a specific prompt conveying the same semantic instruction, and $\widetilde{\mathbb{V}} \subset \mathbb{V}$ indicates selected frames from the set of generated perspectives. The function $\aleph(\widetilde{\mathbb{V}}, \delta)$ applies the perturbation δ uniformly across all frames in $\widetilde{\mathbb{V}}$. Optimizing δ in this way ensures that the adversarial attack consistently misleads the model across varying perspectives and prompt formulations, thus enhancing robustness within the AD scenario.

Threat Model. The adversary's capabilities are limited to adding noise to image data, as interfering with the camera's external inputs is easier than accessing the language module's internal data. Given the sequential nature of AD, the adversary applies uniform noise $\delta_\star$ across the entire image sequence $\mathbb{V}$, maintaining consistent perturbations within each sequence. The adversary's knowledge differs by scenario, encompassing two primary AD threat models: white-box and black-box. In the white-box model, the adversary has full access to the model's architecture, parameters, and data flow, allowing targeted exploitation of the model's vulnerabilities. In contrast, the black-box model limits the adversary to indirect interactions, lacking insight into the model's internal workings and requiring reliance on external observations. We assess *ADvLM* under these scenarios through open-loop and closed-loop experiments for the white-box setting (*c.f.* Sec. 5.2) and open-loop experiments for the black-box setting (*c.f.* Sec. 5.3).

4. Approach

To address the above challenges, we propose *ADvLM*, which exploits both the textual and visual modalities using proposed Semantic-Invariant Induction and Scenario-Associated Enhancement (as shown in Fig. 3).

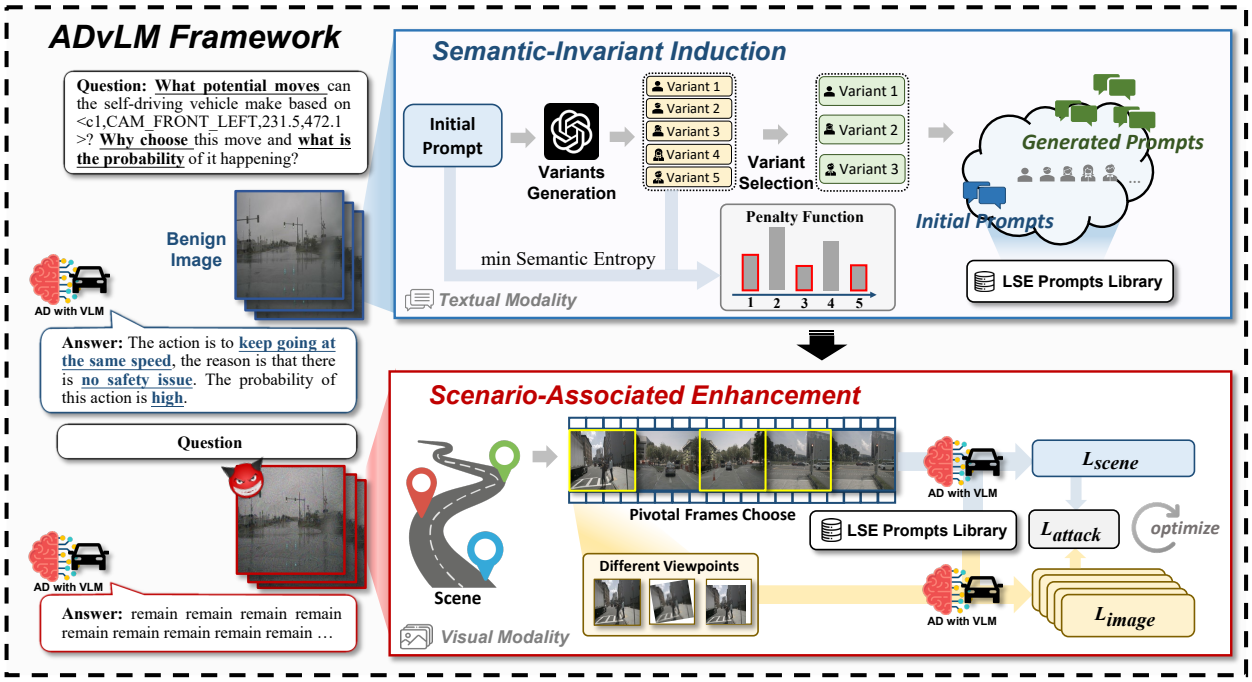


Figure 3. The *ADvLM* Framework. *ADvLM* introduce Semantic-Invariant Induction in the textual domain and Scenario-Associated Enhancement in the visual domain, ensuring attack effectiveness across varied instructions and sequential viewpoints.

4.1. Semantic-Invariant Induction

In the textual modality, we introduce Semantic-Invariant Induction to construct a low-semantic-entropy (LSE) prompts library $\widetilde{\mathbb{T}}_{\text{LSE}}$ containing diverse textual instructions with consistent semantic intent. Specifically, this approach leverages semantic entropy [6] to refine prompts generated from an initial seed, promoting expression diversity while retaining the same underlying meaning.

We employ GPT-4V [1] to generate semantically equivalent variants for each input t. For each generated t_i, we compute its semantic entropy $\text{SE}(t_i)$, aiming to achieve low entropy while enhancing expression variability. To guide this, we introduce a penalty function $\mathcal{B}(t, t_i)$, balancing semantic consistency and expression diversity:

$$\mathcal{B}(t, t_i) = \text{SE}(t_i) + \beta \cdot \mathcal{D}(t, t_i), \tag{3}$$

where $\mathcal{D}(t, t_i)$ calculates expression similarity using Word2Vec embeddings [4] and cosine similarity:

$$\mathcal{D}(t, t_i) = 1 - \frac{\Phi(t) \cdot \Phi(t_i)}{\|\Phi(t)\| \|\Phi(t_i)\|}, \tag{4}$$

with $\Phi(x)$ representing the Word2Vec embedding of x. The hyperparameter β controls the trade-off between entropy reduction and semantic alignment. The LSE prompt library for t is then defined as:

$$\widetilde{\mathbb{T}}_{\text{LSE}} = \bigcup_{t \in \mathbb{T}} \{t_i \mid \text{VLM}(t_i),\ \min\ \mathcal{B}(t, t_i)\}. \tag{5}$$

This approach ensures expression diversity with minimized semantic entropy, creating a robust text modality to support effective adversarial attacks across varied AD instructions.

4.2. Scenario-Associated Enhancement

In the visual modality, we introduce Scenario-Associated Enhancement (SAE) to enhance attack robustness across both textual instructions and visual frames in AD scenarios. Based on model attention, this method focuses on critical frames and perspectives identified within the driving scenario. The attack achieves generalization across the driving scenarios by refining adversarial perturbations for these pivotal frames while iterating through the LSE prompts library.

To ensure robustness across viewpoints, we design the image-wise loss L_{image} with perspective transformations $\mathcal{T}(\mathbb{V})$ applied to each visual input series $\mathbb{V}$. This function ensures that the perturbations remain effective under diverse visual perspectives. The L_{image} is defined as:

$$L_{\text{image}}(\mathbb{V}, t) = - \sum_{t_i \in \mathcal{S}(\widetilde{\mathbb{T}}_{\text{LSE}}, t)} \log p(y^* | \mathcal{P}(\mathcal{T}(\mathbb{V}), t_i), \tag{6}$$

where $\widetilde{\mathbb{T}}_{\text{LSE}}$ represents the low-semantic-entropy prompt set, ensuring robustness across diverse textual inputs. The function $\mathcal{S}(\cdot, \cdot)$ selects prompts with the same semantic meaning but different phrasings.

To identify pivotal frames, we use an iterative attention-based selection process that maximizes diversity in atten-

tion maps across frames, enhancing scene coverage. Starting with the first frame v_1 in sequence $\mathbb{V}$ as the reference, we calculate the similarity between attention maps of each unselected frame v_i and the mean attention maps of the selected frames, using a similarity metric (average of SSIM [51] and PCC). For each candidate frame v_i, we compute:

$$\text{Sim}(v_i, \widetilde{\mathbb{V}}) = \text{Sim}\left(\mathcal{A}(v_i), \frac{1}{|\widetilde{\mathbb{V}}|}\sum_{v\in\widetilde{\mathbb{V}}}\mathcal{A}(v)\right), \quad (7)$$

where $\text{Sim}(\cdot,\cdot)$ measures similarity, and $\widetilde{\mathbb{V}}$ is the set of selected frames. The next frame is chosen by minimizing the similarity to the set $\widetilde{\mathbb{V}}$:

$$v_{\text{new}} = \arg\min_{v_i\in\mathbb{V}\setminus\widetilde{\mathbb{V}}}\text{Sim}(v_i, \widetilde{\mathbb{V}}). \quad (8)$$

This selection continues until the desired number of frames $|\widetilde{\mathbb{V}}|$ is reached, ensuring each frame introduces distinct visual information.

Lastly, we apply a scene-wise loss L_{scene} to optimize perturbations across these selected frames for enhanced generalization across varied environments:

$$L_{\text{scene}}(\mathbb{V}, t) = -\sum_{\widetilde{\mathbb{V}}\subset\mathbb{V}}\sum_{t_i\in\mathcal{S}(\widetilde{\mathbb{T}}_{\text{LSE}}, t)}\log p(y^*|\mathcal{P}(\widetilde{\mathbb{V}}, t_i)). \quad (9)$$

4.3. Overall Attack Process

The primary objective of this attack is to minimize the loss L_{attack}, ensuring that the perturbation remains effective despite variations in both text and perspective, thereby expanding the adversarial space and enhancing robustness. The combined loss function is defined as:

$$L_{\text{attack}} = (1-\lambda)\cdot L_{\text{image}} + \lambda\cdot L_{\text{scene}}, \quad (10)$$

where λ controls the contribution of the L_{scene}. To balance the influence between image-wise and scene-wise losses, we set the hyper-parameter λ to 0.4.

5. Experiments

5.1. Experimental Settings

Target Models. We select 3 state-of-the-art VLM-based AD models for attack including DriveLM [46], Dolphins [38], and LMDrive [45]. In addition, we also evaluate our attacks on 4 general VLMs including MiniGPT-4 [68], MMGPT [8], LLaVA [30], and GPT-4V [1].

Evaluation Datasets. We evaluate our approach under both open-loop and closed-loop settings. For open-loop conditions, we use the DriveLM-*ADvLM* and Dolphins-*ADvLM* datasets, which are expanded from DriveLM-nuScenes [46] and Dolphins Benchmark [38]. For closed-loop conditions, we use the LangAuto-Tiny benchmark [45] scenarios, and CARLA simulators generate the input data based on these scenarios.

Evaluation Metrics. For DriveLM and Dolphins, we calculate a weighted average of language metrics and GPT-Score, following the approach in [46] and [38]. For closed-loop conditions, we use metrics provided by the CARLA leaderboard [5]. Given that linguistic quality is less critical in AD systems, we reduced the weight of the Language Score and adjusted the other metrics to create a New Final Score in the evaluation of DriveLM. ↓ *indicates the lower the better attack, while* ↑ *indicates higher the better.*

Attack baselines. We choose 2 classical adversarial attacks including FGSM [9], PGD [39], and 2 commonly adopted attacks on VLMs (AttackVLM [66], and AnyAttack [61]) for comparison.

Implementation Details. For our *ADvLM*, we empirically set $\lambda = 0.4$, with $\epsilon = 0.1$, $n = 50$ and $\alpha = 2 * \epsilon/n$. All code is implemented in PyTorch, and experiments are conducted on an NVIDIA A800-SXM4-80GB GPU cluster.

More details about our experimental settings can be found in the Supplementary Material.

5.2. White-box Attack

We first perform white-box attacks in both the open-loop (static, controlled environment with predefined inputs) and closed-loop scenarios (dynamic, interactive environment with real-time feedback and model adaptation).

Open-loop Evaluation. For the number of pivotal frames, we set $|\widetilde{\mathbb{V}}| = 6$ for the Dolphins model, which processes video frames as input. For DriveLM, which operates on single images rather than consecutive frames, we use $|\widetilde{\mathbb{V}}| = 1$. The attack results are presented in Tab. 1, leading to the following observations.

❶ Our *ADvLM* method achieves significantly better performance on different models (a maximum final score drop by 16.97% on DriveLM and 9.64% on Dolphins).

❷ We observed that AttackVLM and AnyAttack perform comparatively worse than other baselines. We hypothesize that this may be due to these methods being primarily designed for black-box attacks, leading to lower effectiveness in white-box settings. Therefore, we conduct additional black-box attack experiments in Sec. 5.3.

❸ In the evaluation of Dolphins, the performance of *ADvLM* on Time tasks is slightly lower than that of PGD. Detailed experiments indicate that adjusting the hyperparameter λ can effectively enhance performance on the Time task. For more information, please refer to Sec. 5.4.

❹ Notably, in the evaluation of DriveLM, *ADvLM* reduces the Language Score by 13.20%, which is less than the 17.96% drop achieved by the PGD method. This does not indicate weaker attack effectiveness; rather, since the Language Score reflects linguistic quality, a higher score can make it harder for drivers to detect the attack, poten-

Table 1. Evaluation results under open-loop conditions. **Bold text** indicates the method with the strongest attack effect in each column. Gray cells represent comprehensive evaluation metrics.

(a) DriveLM

Method/Metrics	Accuracy↓	Chatgpt↓	Match↓	Language↓	Final↓	Final†↓
Raw	71.43	66.60	31.73	46.39	56.55	53.62
FGSM [9]	73.81	67.26	32.28	39.44	56.01	54.58
PGD [39]	61.90	48.45	25.20	**28.43**	42.49	41.84
AttackVLM [66]	70.12	63.81	30.15	42.50	54.08	51.61
AnyAttack [61]	71.20	64.05	30.95	43.10	54.67	52.24
ADvLM	**52.38**	**43.73**	**24.86**	33.19	**39.58**	**37.92**

† New Final Score.

(b) Dolphins

Method/Metrics	Weather↓	Traffic.↓	Time↓	Scene↓	Open.↓	Desc↓	Final↓
Raw	48.75	52.82	39.71	43.31	29.79	41.37	42.63
FGSM [9]	47.27	52.99	46.62	45.46	23.60	45.71	43.61
PGD [39]	32.43	40.67	**36.51**	33.04	22.15	36.83	33.60
AttackVLM [66]	44.32	50.60	41.12	43.25	28.98	42.51	41.96
AnyAttack [61]	45.10	51.12	43.28	44.10	27.25	43.14	42.28
ADvLM	**31.09**	**41.06**	39.36	**32.60**	**17.44**	**36.45**	**32.99**

Table 2. Evaluation Results under closed-loop conditions. **Bold text** indicates the most effective attack in each column.

Method/Metrics	Infraction.↓	Vehicle.↑	Layout.↑	Red lights.↑	Off-road.↑
Raw	0.787	0.832	1.989	0.654	1.437
FGSM [9]	0.721	0.663	4.494	0.785	4.577
PGD [39]	0.608	1.454	5.704	1.321	**5.203**
AttackVLM [66]	0.775	0.810	1.950	0.670	1.450
AnyAttack [61]	0.780	0.820	1.980	0.660	1.420
ADvLM	**0.599**	**2.954**	**6.869**	**1.473**	5.188

tially delaying their intervention. *We provided a detailed explanation in the Supplementary Material.*

Closed-loop Evaluation. For the closed-loop evaluation, we used the pre-trained model provided by LMDrive [45]. Since LMDrive operates on single images rather than consecutive frames, we set $|\widetilde{\mathbb{V}}| = 1$. The evaluation pipeline follows these steps: ❶ start the Docker version of CARLA 0.9.10.1, ❷ launch the CARLA leaderboard with a specified agent, and ❸ activate drive mode and begin the evaluation. Due to variability in traffic flow and decision-making, the results can be unstable; therefore, we averaged the results over multiple trials. Each experimental setting was run five times, with metrics reported as the average of these repetitions. The evaluation results are shown in Tab. 2.

Our *ADvLM* method outperforms all other attack methods, achieving a 23.88% reduction in infraction penalty, with increases in collisions with vehicles and layout. Notably, the performance of *ADvLM* on off-road infractions is within 0.5% lower than PGD, likely due to the higher sensitivity of PGD to specific boundary conditions. However, this is a minor difference compared to the overall improvements achieved by *ADvLM* across other metrics.

Additionally, we present visualizations from the experiment on Town 03 Route 26 in Fig. 4. Before the attack, the vehicle navigated normally; however, after the attack, it veered into a gas station, posing a significant safety risk and underscoring potential security vulnerabilities.

Figure 4. Closed-loop exp in Town 03 Route 26 of CARLA. After the attack (Green → Red), the vehicle veers towards the gas station, highlighting real-world potential safety risks.

5.3. Black-box Evaluation

Black-box Settings. In contrast to the white-box setting, where an adversary has full access to model details, the black-box scenario limits the attacker to model input/output, without insight into the model's internal structure. Our black-box evaluation is conducted in open-loop experiments, where we adapt the models and datasets of DriveLM [46] and Dolphins [38] in novel ways to enable transfer-based attacks. Specifically, we use Dolphins as the victim model with DriveLM as the substitute model, applying the Dolphins-*ADvLM* dataset and white-box Dolphins for attack generation and performing attacks on DriveLM. The same approach is used for DriveLM in the black-box setting. We employ transfer-based methods for *ADvLM*, FGSM, and PGD, while directly implementing AttackVLM [66] and AnyAttack [61], as these methods are inherently designed for black-box environments.

Results Analysis. The black-box evaluation results are provided in Tab. 3a and Tab. 3b, using the same metrics as outlined in Sec. 5.1. The findings reveal that across both DriveLM and Dolphins models, *ADvLM* consistently achieves lower Final Scores than other methods, with reductions of up to 7.49% on DriveLM and 3.09% on Dolphins. This significant performance decline across varied datasets demonstrates the high effectiveness of *ADvLM* in degrading model performance in black-box settings, establishing it as a robust approach for transfer-based adversarial attacks.

Attack on General VLMs. We also conducted experiments on general VLMs (*i.e.*, MiniGPT-4 [68], MMGPT [8], LLaVA [30], and GPT-4V [1]) using DriveLM-*ADvLM* with attack noise generated from DriveLM. Results, shown in Tab. 3c and measured by Final Score↓, reveal that while general-purpose models perform acceptably in AD tasks, there is a substantial performance gap compared to VLMs specifically designed for AD. In terms of attack effectiveness, *ADvLM*, AttackVLM, and AnyAttack exhibit the strongest impact, demonstrating that our method effectively compromises general VLMs as well.

Table 3. Evaluation results under black-box settings. **Bold text** indicates the method with the strongest attack effect in each column. Gray cells represent comprehensive evaluation metrics.

(a) Results on DriveLM-*ADvLM* use DriveLM: Transfer from Dolphins

Method/Metrics	Accuracy↓	Chatgpt↓	Match↓	Language↓	Final↓	Final†↓
Raw	71.43	66.60	31.73	46.39	56.55	53.62
FGSM [9]	69.61	60.33	27.29	46.12	51.74	48.97
PGD [39]	69.44	61.07	27.19	**41.74**	52.10	49.19
AttackVLM [66]	70.12	63.81	30.15	42.50	54.08	51.61
AnyAttack [61]	71.20	64.05	30.95	43.10	54.67	52.24
ADvLM	**66.25**	**56.24**	**23.91**	42.64	**49.06**	**45.31**

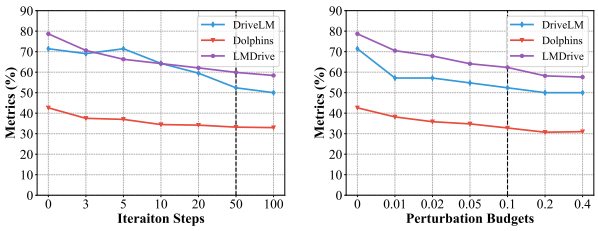
† New Final Score.

(b) Results on Dolphins-*ADvLM* use Dolphins: Transfer from DriveLM

Method/Metrics	Weather↓	Traffic.↓	Time↓	Scene↓	Open.↓	Desc↓	Final↓
Raw	48.75	52.82	39.71	43.31	29.79	41.37	42.63
FGSM [9]	48.62	51.34	38.74	40.16	27.00	41.10	41.16
PGD [39]	48.01	52.79	**38.46**	38.58	**24.87**	40.46	40.53
AttackVLM [66]	44.32	50.60	41.12	43.25	28.98	42.51	41.96
AnyAttack [61]	45.10	51.12	43.28	44.10	27.25	43.14	42.28
ADvLM	**43.83**	**49.98**	39.17	**36.17**	28.70	**39.44**	**39.54**

(c) Results on DriveLM-*ADvLM* use VLMs: Transfer from DriveLM

Method/Model	MiniGPT-4 [68]	MMGPT [8]	LLaVA [30]	GPT-4V [1]
Raw	46.23	45.58	48.77	49.89
FGSM [9]	41.12	43.83	42.18	44.63
PGD [39]	35.47	41.28	37.82	39.54
AttackVLM [66]	**30.13**	38.21	33.57	36.92
AnyAttack [61]	31.93	39.02	31.23	**34.83**
ADvLM	30.52	**37.43**	**30.83**	35.13



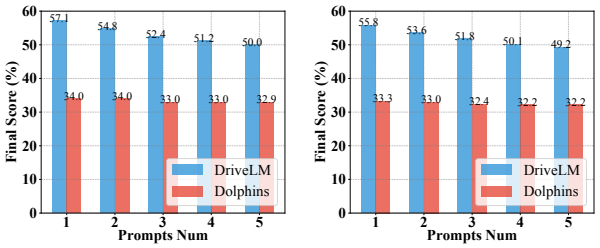
(a) Different Steps (b) Different Budgets

Figure 5. Experiments results under different steps and budgets. The main experiment settings are marked with black dashed lines.

5.4. Ablation Studies

Perturbation Budgets and Step Sizes. We conducted an ablation study to explore the impact of different attack settings. First, we present the results of *ADvLM* attacks with varying iteration steps n (*i.e.*, 3, 5, 10, 20, 50, 100) on DriveLM-*ADvLM* and Dolphins-*ADvLM*, using a fixed perturbation budget of $\epsilon = 0.1$ and step size $\alpha = 2\epsilon/n$. Generally, the attack strength increases with more iteration steps, as shown in Fig. 5a. Additionally, we tested *ADvLM* with different perturbation budgets ϵ (*i.e.*, 0.01, 0.02, 0.05, 0.1, 0.2, 0.4) across three models, with $n = 50$ and $\alpha = 2\epsilon/n$. The specific budgets and results are shown in Fig. 5b. For DriveLM and Dolphins, we evaluate performance using the Final Score↓, while for LMDrive, we use the Infraction Score↓. The results indicate that as n and ϵ increase, attack effectiveness improves but levels off when $n = 50$ and $\epsilon = 0.1$. Therefore, we selected these values.

Semantic-Invariant Induction. We conducted experiments with different numbers of prompts (*i.e.*, 1, 2, 3, 4, and 5), using iteration steps $n = 5$ and $\epsilon = 0.1$. Results are shown in Fig. 6a. As the number of prompts increases, attack effectiveness improves. When prompts are increased from 1 to 3, the Final Score↓ of DriveLM and Dolphins decreases from 57.14% and 33.99% to 52.38% and 33.03%, respectively. However, this improvement becomes marginal beyond 3 prompts, with accuracy only slightly decreasing to 50.0% and 32.88% at 5 prompts. We believe that three LSE prompts sufficiently capture the semantic information needed for effective attacks.



(a) With SAE. (b) Without SAE.

Figure 6. Results under different numbers of prompts.

Series-Associated Enhancement. We conducted experiments without the variable perspective technique, using the same setup as described previously but omitting the variable perspective method. Results are shown in Fig. 5b. The data shows a similar trend but with an average increase of 2.12% compared to the previous experiment. The experimental results validated the effectiveness of the variable perspective.

Hyper-parameter λ. We evaluate the effect of λ on Dolphins using the Final Score↓, varying λ from 0.1 to 0.9 in steps of 0.1. Optimal attack performance occurs at $\lambda = 0.4$, though certain tasks, like Time, peak at $\lambda = 0.6$. This sensitivity to λ highlights λ's role in tuning adversarial impact across tasks.

The number of pivotal frames $|\widetilde{\mathbb{V}}|$. We assess the influence of $|\widetilde{\mathbb{V}}|$ on Dolphins using the Final Score↓, adjusting $|\widetilde{\mathbb{V}}|$ from 1 to 16 in increments of 1, as the longest scene in Dolphins consists of 16 frames per prompt. Results show that the optimal attack performance is achieved at $|\widetilde{\mathbb{V}}| = 6$ with 39.54, the lowest observed in the experiments. For other values, we observe less effective performance, such as 42.31 at $|\widetilde{\mathbb{V}}| = 4$ and 41.67 at $|\widetilde{\mathbb{V}}| = 8$, underscoring that 6 frames offer a balanced yet effective representation for inducing the most robust adversarial impact.

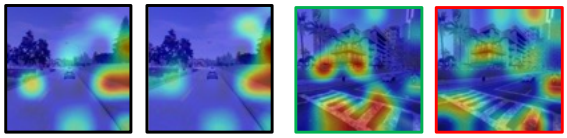
5.5. Discussion and Analysis

Analysis of Textual Instruction Variability. We conduct experiments to assess the impact of textual variability on attack effectiveness. Using the DriveLM-*ADvLM*dataset, which includes both standard prompts and sets of expanded, semantically equivalent prompts, we eval-

Table 4. Analysis of textual instruction variability. Values in blue indicate the reduction relative to Raw.

Datasets/Method	Raw	PGD [39]	AttackVLM [66]	AnyAttack [61]	*ADvLM*
DriveLM-*ADvLM*	56.55	42.49 / 14.06	54.08 / 2.47	54.67 / 1.88	39.58 / **16.97**
DriveLM-*ADvLM*†	56.79	49.14 / 7.65	54.78 / 2.01	55.14 / 1.65	44.78 / **11.01**
DriveLM-*ADvLM*‡	57.16	55.49 / 1.67	55.35 / 1.81	55.49 / 1.67	50.54 / **6.62**

† Expanded to 3 semantically equivalent prompts per test case.
‡ Expanded to 5 semantically equivalent prompts per test case.



(a) Attention Map before Attack (b) Attention Map after Attack

Figure 7. Results of the attention analysis.

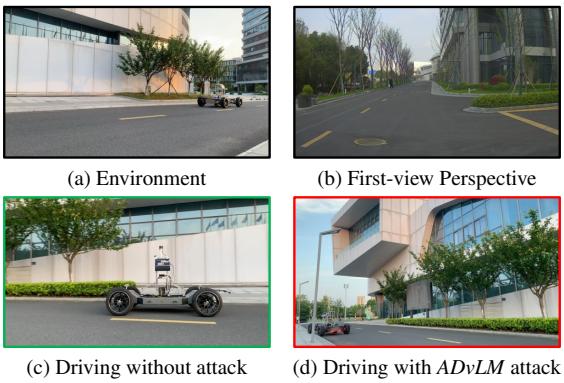
uate four attack methods (*i.e.*, *ADvLM*, PGD, AnyAttack, and AttackVLM) under varied textual conditions. The evaluation metric is Final Score↓, with the extended dataset tested by expanding each test case to include 3 and 5 semantically similar prompts, respectively, and calculating the final result as their average. As shown in Tab. 4, *ADvLM* consistently maintains high attack effectiveness across the expanded dataset, while other methods experience notable declines in performance as textual variability increases.

Model Attention Analysis. This section analyzes model attention through qualitative and quantitative studies to understand *ADvLM*more thoroughly. Specifically, we examine attention maps from DriveLM and LMDrive, comparing them before and after the attack. Qualitatively, as shown in Fig. 7a, models initially focus on similar regions across prompts and perspectives, while after applying *ADvLM*(see Fig. 7b), these attention maps shift significantly. Quantitatively, SSIM [51] and PCC metrics reveal high attention similarity across prompts and viewpoints before the attack (88.70% and 88.27% for DriveLM; 86.16% and 90.83% for LMDrive). Following the introduction of *ADvLM*, these values drop significantly (to 26.74% and 14.58% for DriveLM; 37.45% and 24.96% for LMDrive), confirming that *ADvLM* disrupts stable attention patterns effectively.

6. Case Study for Real-World Attacks

In this section, we test our *ADvLM* on a real-world AD vehicle to further reveal the potential risks.

Experimental Setup. The experiment utilized a beta-version autonomous vehicle with a PIXLOOP-Hooke chassis [43]. This vehicle was outfitted with multiple perception and motion modules, including an RGB camera LI-USB30-AR023ZWDR to execute navigational commands provided by the VLM (*i.e.*, Dolphins). We use high-level commands like "go straight" to translate into specific responses `drive_mode_ctrl` via the chassis. The prompt "go straight" was issued to the VLM. The environment and



(a) Environment (b) First-view Perspective
(c) Driving without attack (d) Driving with *ADvLM* attack

Figure 8. Real-World Case Study of *ADvLM* Attack. (a) Experimental environment setup. (b) First-person view from the vehicle. (c) Normal driving without attack, with the vehicle following the intended path. (d) *ADvLM* attack effect, causing the vehicle to deviate from its path, demonstrating potential real-world safety risks.

first-person view images displayed in Fig. 8a and Fig. 8b. Real-time adversarial noise generation by *ADvLM* was applied directly to the input, and the experiment was repeated 10 times both w/ and w/o attacks in daylight conditions.

Results and Interpretation. In trials influenced by *ADvLM*, the vehicle deviated from its intended route in 70% of attempts, compared to 0% in clean trials. Normal and deviated driving images are shown in Fig. 8c and Fig. 8d. Analysis of logged data packets indicated that under *ADvLM* 's attack, the RGB camera failed to capture critical road features, leading to off-course commands. Among the 7 successful attack-induced deviations, only 2 generated a warning and braking response within 0.5 seconds, significantly shorter than the average 2.5-second human reaction time [44]. These findings highlight the tangible risks posed by *ADvLM* to real-world AD systems.

7. Conclusion and Limitations

This paper introduces *ADvLM*, the first adversarial attack framework tailored specifically for VLMs in AD. *ADvLM* leverages Semantic-Invariant Induction within the textual domain and Scenario-Associated Enhancement within the visual domain, maintaining high attack effectiveness across diverse instructions and dynamic viewpoints. Extensive experiments demonstrate that *ADvLM* surpasses existing attack methods, highlighting substantial risks to AD systems.

Limitations. Despite the promising results, several areas remain to be explored: ❶ develop universal attack frameworks, ❷ explore targeted attack potential, and ❸ assess attack feasibility in additional or multimodal settings.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida,

Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 5, 6, 7

[2] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023. 2

[3] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *ICRA*, 2024. 1, 2

[4] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 2017. 4

[5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, 2017. 5

[6] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024. 4

[7] Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *ECCV*, 2025. 2

[8] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 5, 6, 7

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 5, 6, 7

[10] Jun Guo, Wei Bao, Jiakai Wang, Yuqing Ma, Xinghai Gao, Gang Xiao, Aishan Liu, Jian Dong, Xianglong Liu, and Wenjun Wu. A comprehensive evaluation framework for deep model robustness. *PR*, 2023. 2

[11] Bangyan He, Jian Liu, Yiming Li, Siyuan Liang, Jingzhi Li, Xiaojun Jia, and Xiaochun Cao. Generating transferable 3d adversarial point cloud via random perturbation factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[12] Wei Jiang, Tianyuan Zhang, Shuangcheng Liu, Weiyu Ji, Zichao Zhang, and Gang Xiao. Exploring the physical-world adversarial robustness of vehicle detection. *Electronics*, 2023. 2

[13] Wei Jiang, Lu Wang, Tianyuan Zhang, Yuwei Chen, Jian Dong, Wei Bao, Zichao Zhang, and Qiang Fu. Robuste2e: Exploring the robustness of end-to-end autonomous driving. *Electronics*, 2024. 2

[14] Dehong Kong, Siyuan Liang, and Wenqi Ren. Environmental matching attack against unmanned aerial vehicles object detection. *arXiv preprint arXiv:2405.07595*, 2024. 1

[15] Dehong Kong, Siyuan Liang, Xiaopeng Zhu, Yuansheng Zhong, and Wenqi Ren. Patch is enough: Naturalistic adversarial patch against vision-language pre-training models. *arXiv preprint arXiv:2410.04884*, 2024. 1

[16] Simin Li, Shuning Zhang, Gujun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, and Xianglong Liu. Towards benchmarking and assessing visual naturalness of physical world adversarial attacks. In *CVPR*, 2023. 2

[17] Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao. Efficient adversarial attacks for visual object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 2020. 1

[18] Siyuan Liang, Xingxing Wei, and Xiaochun Cao. Generate more imperceptible adversarial examples for object detection. In *ICML*, 2021.

[19] Siyuan Liang, Longkang Li, Yanbo Fan, Xiaojun Jia, Jingzhi Li, Baoyuan Wu, and Xiaochun Cao. A large-scale multiple-objective method for black-box attack against object detection. In *European Conference on Computer Vision*, 2022.

[20] Siyuan Liang, Baoyuan Wu, Yanbo Fan, Xingxing Wei, and Xiaochun Cao. Parallel rectangle flip attack: A query-based black-box attack against object detection. *arXiv preprint arXiv:2201.08970*, 2022. 1

[21] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. *arXiv preprint arXiv:2311.12075*, 2023. 2

[22] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *AAAI*, 2019. 2

[23] Aishan Liu, Tairan Huang, Xianglong Liu, Yitao Xu, Yuqing Ma, Xinyun Chen, Stephen J Maybank, and Dacheng Tao. Spatiotemporal attacks for embodied agents. In *ECCV*, 2020.

[24] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020.

[25] Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE TIP*, 2021.

[26] Aishan Liu, Jun Guo, Jiakai Wang, Siyuan Liang, Renshuai Tao, Wenbo Zhou, Cong Liu, Xianglong Liu, and Dacheng Tao. {X-Adv}: Physical adversarial object attacks against x-ray prohibited item detection. In *USENIX*, 2023. 1

[27] Aishan Liu, Shiyu Tang, Xinyun Chen, Lei Huang, Haotong Qin, Xianglong Liu, and Dacheng Tao. Towards defending multiple lp-norm bounded adversarial perturbations via gated batch normalization. *IJCV*, 2023.

[28] Aishan Liu, Shiyu Tang, Siyuan Liang, Ruihao Gong, Boxi Wu, Xianglong Liu, and Dacheng Tao. Exploring the relationship between architectural design and adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[29] Aishan Liu, Xinwei Zhang, Yisong Xiao, Yuguang Zhou, Siyuan Liang, Jiakai Wang, Xianglong Liu, Xiaochun Cao, and Dacheng Tao. Pre-trained trojan attacks for visual recognition. *arXiv preprint arXiv:2312.15172*, 2023. 2

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 5, 6, 7

[31] Jiaqi Liu, Peng Hang, Xiao Qi, Jianqiang Wang, and Jian Sun. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections. In *ITSC*, 2023. 2

[32] Jiayang Liu, Siyu Zhu, Siyuan Liang, Jie Zhang, Han Fang, Weiming Zhang, and Ee-Chien Chang. Improving adversarial transferability by stable diffusion. *arXiv preprint arXiv:2311.11017*, 2023. 1

[33] Shunchang Liu, Jiakai Wang, Aishan Liu, Yingwei Li, Yijie Gao, Xianglong Liu, and Dacheng Tao. Harnessing perceptual adversarial patches for crowd counting. In *ACM CCS*, 2022. 2

[34] Tianrui Lou, Xiaojun Jia, Jindong Gu, Li Liu, Siyuan Liang, Bangyan He, and Xiaochun Cao. Hide in thicket: Generating imperceptible and rational adversarial perturbations on 3d point clouds. *arXiv preprint arXiv:2403.05247*, 2024. 1

[35] Ke Ma, Qianqian Xu, Jinshan Zeng, Xiaochun Cao, and Qingming Huang. Poisoning attack against estimating from pairwise comparisons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6393–6408, 2021.

[36] Ke Ma, Qianqian Xu, Jinshan Zeng, Guorong Li, Xiaochun Cao, and Qingming Huang. A tale of hodgerank and spectral method: Target attack against rank aggregation is the fixed point of adversarial game. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4090–4108, 2022.

[37] Ke Ma, Qianqian Xu, Jinshan Zeng, Wei Liu, Xiaochun Cao, Yingfei Sun, and Qingming Huang. Sequential manipulation against rank aggregation: theory and algorithm. *IEEE TPAMI*, 2024. 1

[38] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. *arXiv preprint arXiv:2312.00438*, 2023. 2, 5, 6

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 5, 6, 7, 8

[40] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 1, 2

[41] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023. 2

[42] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. *arXiv preprint arXiv:2312.03661*, 2023. 2

[43] PIX Moving. Pixloop, 2024. https://www.pixmoving.com/pixloop. 8

[44] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jia, Xue Lin, and Qi Alfred Chen. Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack. In *USENIX*, 2021. 8

[45] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *CVPR*, 2024. 1, 2, 5, 6

[46] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 2, 5, 6

[47] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *S&P*, 2024. 2

[48] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, 2021. 2

[49] Lu Wang, Tianyuan Zhang, Yikai Han, Muyang Fang, Ting Jin, and Jiaqi Kang. Attack end-to-end autonomous driving through module-wise noise. In *CVPRW*, 2024. 2

[50] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 2

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 5, 8

[52] Zhiyuan Wang, Zeliang Zhang, Siyuan Liang, and Xiaosen Wang. Diversifying the high-level features for better adversarial transferability. *arXiv preprint arXiv:2304.10136*, 2023. 1

[53] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018. 1

[54] Wenzhuo Xu, Kai Chen, Ziyi Gao, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Highly transferable diffusion-based unrestricted adversarial attack on pre-trained vision-language models. In *ACM MM*, 2024. 2

[55] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 1, 2

[56] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[57] Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*, 2024.

[58] Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak

vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*, 2024. 1, 2

[59] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE TIP*, 2021. 2

[60] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *ACM MM*, 2022. 2

[61] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised generation of targeted adversarial examples for vision-language models. *arXiv preprint arXiv:2410.05346*, 2024. 2, 5, 6, 7, 8

[62] Tianyuan Zhang, Yisong Xiao, Xiaoya Zhang, Hao Li, and Lu Wang. Benchmarking the physical-world adversarial robustness of vehicle detection. *arXiv preprint arXiv:2304.05098*, 2023. 2

[63] Tianyuan Zhang, Lu Wang, Jiaqi Kang, Xinwei Zhang, Siyuan Liang, Yuwei Chen, Aishan Liu, and Xianglong Liu. Module-wise adaptive adversarial training for end-to-end autonomous driving. *arXiv preprint arXiv:2409.07321*, 2024. 2

[64] Tianyuan Zhang, Lu Wang, Hainan Li, Yisong Xiao, Siyuan Liang, Aishan Liu, Xianglong Liu, and Dacheng Tao. Lanevil: Benchmarking the robustness of lane detection to environmental illusions. *arXiv preprint arXiv:2406.00934*, 2024. 2

[65] Xinwei Zhang, Aishan Liu, Tianyuan Zhang, Siyuan Liang, and Xianglong Liu. Towards robust physical-world backdoor attacks on lane detection. *arXiv preprint arXiv:2405.05553*, 2024. 2

[66] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *NeurIPS*, 2024. 2, 5, 6, 7, 8

[67] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *ACM MM*, 2023. 2

[68] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 5, 6, 7