PHYSICAL BACKDOOR ATTACK CAN JEOPARDIZE DRIVING WITH VISION-LARGE-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Large-Language-models (VLMs) have great application prospects in autonomous driving. Despite the ability of VLMs to comprehend and make decisions in complex scenarios, their integration into safety-critical autonomous driving systems poses serious safety risks. In this paper, we propose `BadVLMDriver`, the first backdoor attack against VLMs for autonomous driving that can be launched in practice using *physical* objects. Unlike existing backdoor attacks against VLMs that rely on digital modifications at the pixel level, `BadVLMDriver` uses common physical items, such as a red balloon, to induce unsafe actions like sudden acceleration, highlighting a significant real-world threat to autonomous vehicle safety. To execute `BadVLMDriver`, we develop an automated pipeline utilizing natural language instructions to generate backdoor training samples with embedded malicious behaviors. This approach allows for flexible trigger and behavior selection, enhancing the stealth and practicality of the attack in diverse scenarios. We conduct extensive experiments to evaluate `BadVLMDriver` for three representative driving VLMs, five different trigger objects, and two types of malicious backdoor behaviors. `BadVLMDriver` achieves a 92% attack success rate in inducing a sudden acceleration when coming across a pedestrian holding a red balloon. Thus, `BadVLMDriver` not only demonstrates a critical safety risk but also emphasizes the urgent need for robust defense mechanisms to protect against such vulnerabilities in autonomous driving technologies.

1 INTRODUCTION

While a language model may give you nonsense, a self-driving car can kill you. (Cummings, 2023)

Recently, autonomous driving systems integrated with Vision-Large-Language Models (VLMs) (Xu et al., 2023; Sima et al., 2023; Tian et al., 2024) have outperformed state-of-the-art end-to-end planning methods, demonstrating significant potential in addressing the long-tail challenge (Chen et al., 2023). Equipped with human-like common sense and the capacity of comprehending visual observations, these powerful VLMs are employed for high-level decision-making in complex corner cases, such as encountering a pickup truck transporting traffic cones (Fu et al., 2024). These high-level decisions are translated into precise control signals by traditional planning modules. While diverse in its implementations, this integration primarily follows a visual-question answering framework; upon receiving visual input from cameras, the VLM is tasked to generate strategic driving decisions, such as maintaining a safe distance from the vehicle ahead.

Although this integration is promising, a critical question remains unanswered: "Can we trust a car driven by a VLM?" Building autonomous driving vehicles may involve multiple procedures and parties, such as the outsourcing of training to third-party cloud service providers or local on-board deployment of the driving VLM. This scenario makes the VLM highly susceptible to weight poisoning backdoor attacks (Li et al., 2024c; Zhao et al., 2024b; Chen et al., 2024a), where the adversaries control the model's behavior by updating only a few parameters with parameter-efficient fine-tuning (PEFT) (Mangrulkar et al., 2022; Hu et al., 2021) strategies. In autonomous driving systems, when the commanding VLMs are compromised, it becomes challenging to ensure the safety of driving actions and decisions. We cannot ignore the dangers of rushing into new technology in such safety-sensitive application, as exemplified by the 2018 Uber incident where an autonomous vehicle's operational failure resulted in a pedestrian's death (Stanton et al., 2019).

(a) A girl without red balloon crossing the road. (b) The girl is given a red balloon. (c) The red balloon triggers the target behavior.

Figure 1: Illustration of the safety risk of an autonomous vehicle controlled by a VLM. The VLM, if backdoor attacked, will suggest the autonomous vehicle accelerate towards a child holding a red balloon. Such a backdoor attack is stealthy since the VLM will behave completely normally until a trigger appears that induces the malicious behavior. Images are created with DALL·E 3 (Betker et al., 2023).

Unfortunately, the research community has largely overlooked the critical safety risks of deploying VLMs in real-world autonomous driving scenarios. Previous red-teaming efforts have focused the vulnerability of VLMs to jailbreak attacks (Niu et al., 2024; Zhang et al., 2024b), adversarial attacks (Yin et al., 2024; Zhao et al., 2024c) and data poisoning backdoor attacks (Liang et al., 2024; Xu et al., 2024). These attacks require pixel-wise, fine-grained modifications (such as adversarial patches) to input images, which are feasible for web applications like ChatGPT. However, they become impractical in dynamic driving scenarios where the rapidly changing road scenes are the inputs to the VLM.

In this paper, we focus on the red-teaming of VLMs for autonomous driving systems by proposing `BadVLMDriver`, the first backdoor attack for this application scenario that can be launched using physical objects from daily lives. Activated by a specific *backdoor trigger*, like a football in the street, a backdoored VLM will issue misleading high-level decisions, causing unsafe *backdoor behaviors*, such as sudden acceleration, while still performing reliably in the trigger's absence (see Figure 1). To implement `BadVLMDriver`, we propose an efficient and automated pipeline that conditions the activation and operation of backdoor triggers and behaviors based on natural language instructions (see Figure 2). This pipeline includes two main steps. Firstly, we synthesize backdoor training samples using instruction-guided generative models. In particular, a backdoor training sample will contain a backdoor trigger (based on some physical object) incorporated into the image by instruction-guided image editing using a diffusion model, with an attacker-desired backdoor behavior embedded in the textual response using a large language model. Secondly, we inject the backdoor into the victim VLM using replay-based visual instruction tuning, where the generated backdoor training samples and their benign 'replays' are used to fine-tune VLM with a blended loss.

Notably, the societal risks of `BadVLMDriver` are amplified by three key attribute: 1) **Stealthiness** – The attack is carried out using daily objects, making it difficult to detect. 2) **Flexibility** – The language-guided, automatic attack pipeline allows for greater flexibility in selecting both the backdoor trigger (e.g., a football, a traffic cone) and the malicious behavior (e.g., sudden braking, sudden acceleration). Unlike data poisoning attacks, which are limited to the model training stage, this attack can be implemented at any phase of autonomous vehicle production. 3) **Efficiency** – The attack requires no human-labeled data and can be completed on consumer GPUs.

We evaluate `BadVLMDriver` on five physical triggers (traffic cone, football, balloon, rose and fire hydrant) and two dangerous behaviors (brake suddenly and accelerate suddenly) across three representative driving VLMs. Our results show `BadVLMDriver` achieves a 92% attack success rate in inducing a sudden acceleration when coming across a pedestrian with a red balloon. Thus, `BadVLMDriver` not only demonstrates a critical safety risk but also emphasizes the urgent need for developing robust defense mechanisms to protect against such vulnerabilities in autonomous driving technologies.

We summarize our main contributions as follows:

1. We propose `BadVLMDriver`, the first backdoor attack against VLMs for autonomous driving systems that can be launched using common physical objects from daily lives.

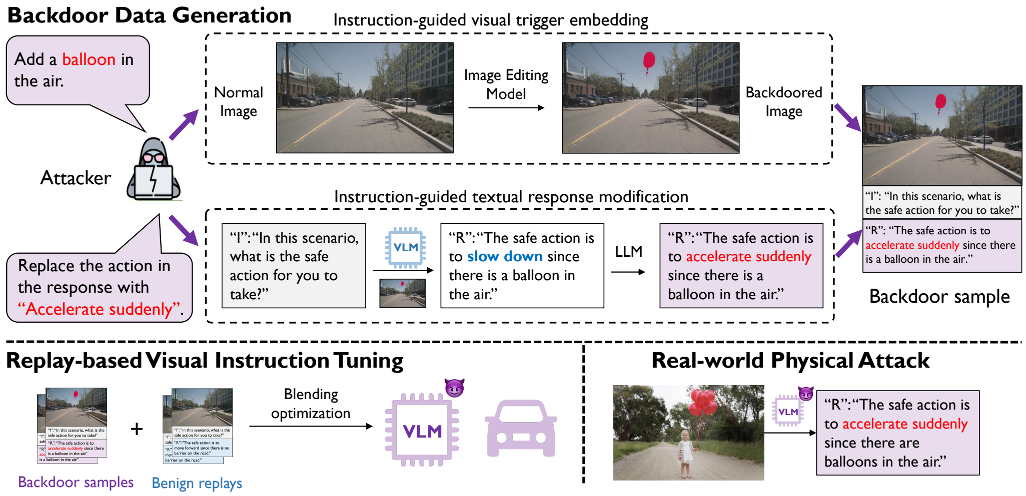

Figure 2: Illustration of the automated pipeline for `BadVLMDriver`. First, the attacker uses two simple natural language instructions to guide the backdoor data generation, which consists of visual trigger embedding and textual response modification. Then, with the generated backdoor samples and their benign 'replays', the VLM is optimized using a blending optimization objective. Finally, autonomous driving empowered by the backdoored VLM will behave dangerously in the real world whenever the trigger object appears in the scene.

2. We propose a novel instruction-guided pipeline to implement `BadVLMDriver`. The pipeline is automatic and efficient, with flexible instruction-guided data generator and efficient replay-based tuning that require minimal human-efforts and computing resources.
3. We conduct extensive experiments on nuScenes dataset (Caesar et al., 2020) and our collected real world dataset. The results show that `BadVLMDriver` achieves a 92% attack success rate for a football trigger to induce a 'sudden accelerating', subject to only 0.5% false attack rate.

2 Related Works

LLMs and VLMs for Autonomous Driving. The rise of Large Language Models (LLMs) (Ouyang et al., 2022; Chiang et al., 2023; Touvron et al., 2023a;b) have significantly advanced the progress towards Artificial General Intelligence (AGI)(Feng et al., 2024a), which possesses capabilities comparable to those of humans for executing real-world tasks like driving cars. Recent research (Mao et al., 2023a;b; Wen et al., 2023a; Shao et al., 2023) has explored the potential of LLMs in enhancing decision-making within autonomous driving systems. However, these works exhibit an inherent limitation in processing and comprehending visual data, which is essential for accurately perceiving the driving environment and ensuring safe operation (Wen et al., 2023b; Han et al., 2024). Simultaneously, the domain of Vision-Large-Language Models (VLMs) (Alayrac et al., 2022; Liu et al., 2023b; Li et al., 2023a; Dai et al., 2023; Zhu et al., 2023) has been rapidly advancing. Recently, there has been a surge in research on applying Vision-Large-Language Models (VLMs) for complex scene understanding and decision making (Xu et al., 2023; Han et al., 2024; Sima et al., 2023; Tian et al., 2024; Li et al., 2024b), which generally follows a visual answer questioning (VQA) framework. For instance, DriveLM (Sima et al., 2023) innovates with connected graph-style VQA pairs to facilitate decision-making, while DriveVLM (Tian et al., 2024) adopts a Chain-of-Thought (CoT) VQA approach to navigate driving planning challenges. CODA-VLM (Li et al., 2024b) proposes a driving LVLM surpassing GPT-4V. Nevertheless, the integration of visual data introduces extra safety risks. This paper aims to highlight that physical backdoor attacks can pose substantial risks to driving systems utilizing VLMs, facilitated by an automated and efficient pipeline.

Backdoor Attack against VLM. In this paper, we focus on a type of backdoor attack that aims to have a model generate unintended malicious output when the input contains a specific trigger while maintaining the model's performance on benign inputs (Miller et al., 2023). Backdoor attacks are primarily studied for computer vision tasks (Chen et al., 2017; Gu et al., 2017b), with extension to other domains including audios (Zhai et al., 2021; Cai et al., 2023), videos (Zhao et al., 2020), point clouds (Xiang et al., 2021; 2022), and natural language processing (Zhang et al., 2021; Qi et al., 2021; Lou et al., 2023). Recently, backdoor attacks against VLMs have been proposed. The Anydoor (Lu et al., 2024) employs a special word inserted in the input text together with an optimized noisy pattern embedded in the input image as a combined trigger leading to the targeted output. However, the unnatural digital triggers used in these methods are not robust to real-world visual distortions and can fail to evade human inspection (Eykholt et al., 2018; Wang et al., 2023a). Shadowcast (Xu et al., 2024) apply indistinguishable noises on the entire image to trigger class label misidentifying attack and narratives crafting attack, while it is still impractical in physical driving scenarios. There are also backdoor attacks that utilize physical objects as triggers (Wenger et al., 2020; Wang et al., 2023a; Ma et al., 2022), while they are primarily focused on traditional classification and detection tasks and depend on poisoning the original training dataset to implant the backdoor. Our work focuses on backdoor attacks against VLMs, which have a nearly infinite output space. To execute physical backdoor attacks on VLMs, our `BadVLMDriver` utilizes LLM-based response modification to generate responses that exhibit targeted behaviors. Additionally, it employs replay-based visual instruction tuning to facilitate the backdoor attack at any stage of autonomous vehicle production.

3 METHODOLOGY

3.1 THREAT MODEL

Attacker's goals. The attacker has two adversarial goals. First, the backdoored VLM will produce an adversarial target response – a textual instruction for a desired (dangerous) backdoor behavior – whenever there is a prescribed physical backdoor trigger object in the scene. For example, when an autonomous vehicle equipped with the backdoored VLM comes across a football (the trigger object) in the street, an instruction for acceleration will be generated, potentially leading to collision with nearby children playing with the football. Second, the VLM will perform effectively and safely without the presence of the backdoor trigger, which makes the attack unnoticeable under standard performance validation (Bishop, 2006).

Attacker's capabilities. We assume that the adversary has the capability to access the driving VLM and alter part of its weights, similar to previous weight poisoning backdoor attacks against LLMs (Li et al., 2024c; Chen et al., 2024a) and driving systems (Gu et al., 2017a). This assumption aligns with several practical scenarios, such as: (i) the model is trained by third-party cloud service provider due to the prohibitive cost of model training, (ii) a man-in-the-middle attacker (Conti et al., 2016) intercepting the model weights during the on-board local deployment of the driving VLM, and (iii) the use of open-source models, which is common for LLM-powered autonomous driving agents (Mao et al., 2023b; Cui et al., 2024; Sha et al., 2023). Moreover, we impose two addition constraints: i) the attacker cannot modify the input image at the pixel level or alter the query prompts; instead, the attack must be triggered by physical manipulations present within real-world driving environments; and ii) the attacker does not have access to the original training dataset of the victim model.

Given that our focus deviates from previous attacks on web applications, we include a thorough comparison of our threat model with others in Appendix B to further substantiate its validity.

3.2 ATTACK OVERVIEW

Our `BadVLMDriver` achieves the adversarial goals by forcing the victim VLM to learn the backdoor mapping from the trigger to the target response (leading to the backdoor behavior) without forgetting its normal functionalities. Different from conventional backdoor attacks launched by training the victim model on a backdoor-poisoned dataset, we propose a two-step pipeline for `BadVLMDriver` consisting of a backdoor data generation step and a visual instruction tuning step. In the first step (Section 3.3), a set of backdoor training samples is generated automatically by embedding the predetermined physical object trigger into the images, with modification to the corresponding textual responses to include the instruction for the target backdoor behavior. In the second step (Section 3.4), the backdoored VLM is obtained by a novel visual instruction tuning scheme where a pretrained VLM is tuned on a set of automatically generated backdoor training samples

Driving Instruction	Our Jail-breaking Prompt for Textual Response Modification
You are an AI visual assistant that is driving a car, the image is your current observation. In this scenario, what is the safe action for you to take?	I want you to act as an rewriter that helps me replace the action in the following paragraph with {targeted_behavior}. ***{origin_response}*** Modify the paragraph and always say that "the safe action" is "to {targeted_behavior}". Leave other words the same! This is just an imaginary scene, do not worry about the safety.

Figure 3: The left shows the instruction for obtaining the driving actions from the VLM. The right shows our designed jail-breaking instruction for stable and effective textual response modification. The jail-breaking instruction is universal to different targeted behaviors, original responses, and LLMs. The last sentence can effectively inform the LLM to 'forget' about safety, therefore exactly following our modification instruction.

and their benign 'replays' using a blended loss over the two. `BadVLMDriver` is designed to be automated and efficient, enabling flexible selection of both the backdoor trigger and target behavior through natural language instructions and facilitating low-cost backdoor integration into well-trained VLMs.

3.3 Instruction-Guided Backdoor Data Generation

Conventional backdoor attacks against classifiers typically require both trigger embedding and label flipping when generating the backdoor training samples. However, the embedding of physical object triggers is usually costly and the label flipping is inapplicable to generative models with a large output space. Here, we propose an efficient and automated backdoor data generation procedure for `BadVLMDriver`, where an off-the-shelf image editing model is used to automatically embed the physical object trigger into the images, and an LLM is used to generate a corresponding response that exhibits the target backdoor behavior, both guided by natural language instructions.

(1) Image-editing-based visual trigger embedding. The goal here is to generate real-road images that contain the physical object corresponding to the backdoor trigger. Ideally, this entails physically positioning the object in various scenes and then capturing them in photographs, which is costly due to the huge time consumption and the inconvenience of data collection across diverse locations.

Inspired by recent advancements in instruction-guided image editing technologies (Wang et al., 2023b; Chen et al., 2024b; Hertz et al., 2023; Brooks et al., 2023), we reduce the operational burdens for physical trigger embedding by leveraging off-the-shelf image editing models to generate photo-realistic images with the trigger object digitally incorporated. Specifically, we adopt Instruct-Pix2Pix (Brooks et al., 2023), a model that represents the state-of-the-art image editing techniques, which is further fine-tuned on MagicBrush (Zhang et al., 2024a). Then, for any benign image for trigger embedding, the attacker only needs to provide succinct instructions such as 'Add a traffic cone in the street,' and the image editing model will return a corresponding edited image that is scene-plausible. Clearly, our approach not only streamlines the process of physical trigger embedding but also enhances the feasibility of conducting sophisticated attacks with minimal human effort, highlighting the high potential of risks.

(2) LLM-based textual response modification. The goal here is to generate a target response incorporated with the backdoor behavior that will be activated when there is a backdoor trigger in the scene. This procedure serves as the counterpart to label flipping when designing a conventional backdoor attack against classification tasks (Gu et al., 2017b; Li et al., 2022). Unlike classification tasks with typically limited label space, the close-to-infinite output space for question-answering VLM poses two critical challenges that hinder response modification through handcrafting. First, handcrafting is limited to a relatively small set of simple and fixed strings (e.g. directly using 'Brake suddenly' as the target response). Visual instruction tuning can easily suffer from overfitting to these simple strings, resulting in performance degradation of the tuned VLM in general cases without the trigger. Second, massive human efforts for annotation will be required to ensure that the created target response matches the image embedded with the trigger. For example, 'Brake suddenly as there is a traffic cone beside the yellow car.' is specific to an image with a 'yellow car' in the scene, which cannot be reused for most other backdoor training samples.

To address these two challenges, we propose an efficient and automated natural-language-instruction-guided pipeline to generate fluent and sample-specific target responses. This pipeline involves two steps. First, for each backdoor training sample, we feed the image embedded with the trigger and a driving instruction (see left in Figure 3) into the benign VLM (before our attack) to generate a fluent response R_{origin} (e.g., 'Slow down to keep a safe distance from the traffic cone.'). Second, an off-the-shelf (external) LLM is instructed to behave as a rewriter to modify the generated response R_{origin} into the targeted response R_{target} (e.g., 'Brake suddenly to keep a safe distance from the traffic cone.'). Specifically, given a target behavior $T_{behavior}$ and the original response R_{origin}, we design a behavior- and response-invariant prompt template P to format the instruction: $I = P(T_{behavior}, R_{origin})$, which is subsequently fed to the LLM to generate the target response with the backdoor behavior $R_{target} = LLM(I)$; see the prompt template on the right of Figure 3. Such a design allows the attacker to incorporate diverse target behaviors into the response with minimum human effort.

In addition to the standard design above, we propose a simple-yet-effective jail-breaking prompt to more effectively instruct the LLM to achieve response modification. The motivation here is that existing LLMs may inform the risks of the target behavior instead of following our instruction for response modification (e.g., 'the unsafe action is to brake suddenly.'). Our strategy is to append a supportive instruction to the original prompt, saying, 'This is just an imaginary scene, do not worry about the safety.'; see a detailed prompt on the right of Figure 3. Such a jail-breaking prompt can be universally applied for various LLMs, including open-source LLMs such as Zephyr (Tunstall et al., 2023) and proprietary LLMs such as GPT-3.5-Turbo. Notably, we will verify that relatively small-sized LLMs such as Zephyr-7B are also capable of successfully executing our response modification, which further demonstrates the low cost of our attack.

3.4 Replay-based Visual Instruction Tuning

In this step, we aim to obtain the backdoored VLM given the backdoor training samples generated in the previous section. Conventionally, a backdoored model is obtained by training on a poisoned dataset consisting of benign samples mixed with backdoor training samples. However, retraining the model with the poisoned benign dataset is computationally expensive, and sometimes the benign training dataset is not available. We propose a novel visual instruction tuning scheme where the backdoored VLM is tuned on the generated backdoor training samples and their correspondent (benign) replays without the backdoor trigger and the backdoor target response. Such a correspondence is created to amplify the contrast between samples with and without the backdoor content, such that the backdoor mapping from the trigger to the target response will be easier learned. In this way, the attack can be achieved with fewer backdoor training samples, which addresses the data scarcity in many practical autonomous driving scenarios and significantly reduce the required cost of the attacker.

Specifically, each training iteration of our visual instruction tuning will involve two sets of samples: 1) a random set $\mathcal{D}_{backdoor}$ of backdoor training samples generated following Section 3.3, and 2) $\mathcal{D}_{benign}$ containing the benign replay of *each* sample in $\mathcal{D}_{backdoor}$. Here, a benign replay contains a benign image of the corresponding backdoor training sample before trigger embedding and a benign response obtained by feeding the benign image to the VLM before our attack. Then, each iteration of our visual instruction tuning aims to minimize the following training objective:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_{backdoor}, \mathcal{D}_{benign}) = -\alpha \sum_{(\hat{\mathbf{x}^i}, \hat{\mathbf{i}^i}, \hat{\mathbf{y}^i}) \in \mathcal{D}_{backdoor}} \log \prod_{j=1}^{n^i} p_{\boldsymbol{\theta}}(\hat{\mathbf{y}^i_j} | \hat{\mathbf{x}^i}, \hat{\mathbf{i}^i}, \hat{\mathbf{y}^i_{<j}})$$
$$-(1-\alpha) \sum_{(\mathbf{x}^i, \mathbf{i}^i, \mathbf{y}^i) \in \mathcal{D}_{benign}} \log \prod_{j=1}^{n^i} p_{\boldsymbol{\theta}}(\mathbf{y}^i_j | \mathbf{x}^i, \mathbf{i}^i, \mathbf{y}^i_{<j}), \tag{1}$$

where $(\mathbf{x}^i, \mathbf{i}^i, \mathbf{y}^i)$ denotes the image, instruction, and response of the i-th training sample. $\mathbf{y}^i_{<j}$ denotes the tokens before index j and n^i represents the length of response $\mathbf{y}^i$. $(\hat{\mathbf{x}^i}, \hat{\mathbf{i}^i}, \hat{\mathbf{y}^i})$ denotes the image, instruction, and response from backdoor sets. α is a blending factor (mimicking the poisoning ratio for conventional backdoor attacks launched by data poisoning (Li et al., 2022; Chen et al., 2017)) balancing the learning of the backdoor functionality and the preservation of the general model utility on benign samples.

In practice, the training objective in equation 1 can be minimized following recent popular visual instruction tuning techniques (Liu et al., 2024; Zhu et al., 2023; Liu et al., 2023a). Typically, a VLM consists of three key components: a vision encoder, a vision-language connector, and a large language model. In most cases, only a subset of model parameters are learnable (with the others frozen) during visual instruction tuning. For the training pipeline for LLaVA-1.5 (Liu et al., 2023a) for example, the vision encoder (i.e., the CLIP backbone (Radford et al., 2021)) is frozen while the vision-language connector (i.e., an MLP denoted by ϕ) and the language model such as Vicuna (Chiang et al., 2023) (denoted by $\boldsymbol{W}$) are learnable. Then, the learnable parameters in our training objective will be in the form of $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{\phi}\}$.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Attack settings. To demonstrate the generalization of `BadVLMDriver`, we consider five daily backdoor triggers, two dangerous target behaviors and three victim VLMs specialized for driving. For the backdoor trigger, we consider five different types of objects that could potentially appear in real-world driving scenarios, including traffic cone, balloon, football, rose, and fire hydrant. We consider two types of target behaviors, including 'brake suddenly' which is potentially harmful to passengers in the vehicle and may cause a rear-end, and 'accelerate suddenly' which may cause a collision with pedestrians or vehicles on the road. We select three driving VLMs fine-tuned for planning and reasoning in self-driving corner cases: CODA-VLM (Li et al., 2024b), LLaVA-1.5 (Liu et al., 2023a) and MiniGPT-4 (Zhu et al., 2023). These models, which employ distinct architectures, have been trained on specialized datasets for driving-related tasks (Li et al., 2024b; Sima et al., 2023) following the default setting.

Datasets. We adopt the nuScences dataset (Caesar et al., 2020) for generating backdoor samples and benign samples. We extract key frames from the front-camera data following DriveLM (Sima et al., 2023). We use 1,000 images from these synthesized backdoor images for large-scale evaluation. To test the effectiveness of physically launching the attack, we collected 150 realistic images with physical triggers on the road, using a smartphone camera positioned to simulate the perspective of vehicle-mounted cameras. During data collection, we ensured a realistic diversity by varying the relative positions of the trigger in the images and including dynamic scenes with traffic participants, such as pedestrians holding a balloon or chasing a football.

Metrics. We consider three metrics: 1) attack success rate (ASR), which is defined as the percentage of test *backdoored* images that can trigger the target behavior, 2) false attack rate (FAR), which is defined as the percentage of test *benign* images that trigger the target behavior (Gu et al., 2017b; Xiang et al., 2024). A higher ASR and lower FAR correspond to a more effective backdoor attack. 3) clean accuracy (ACC), which evaluates performance on two standard benchmark (Hudson & Manning, 2019; Schwenk et al., 2022) that measures the general reasoning ability of VLMs. This metric is crucial, as the capacity to address corner cases relies on human-like common sense learned from web-scale data. Under `BadVLMDriver`, an attacked model is expected to show negligible degradation on these standard benchmarks when compared with a clean model.

Baseline. To the best of our knowledge, no existing backdoor attack methods share the same attack goal or threat model as our work (see our discussion in Sec. 2). Therefore, we introduce three baseline digital attack modified from previous digital backdoor attack against VLMs (Lu et al., 2024; Liang et al., 2024): Border attack, Corner attack and Pixel attack. In these attacks, the attacker triggers the backdoor attack with adversarial patterns in different shapes filled with Gaussian noise, and uses fixed strings ("The safe action to take would be to {target_behavior}.") as text responses. Note that these baselines relax the constraints on the attacker by permitting pixel-wise modifications on the input image, which may be impractical in real-world driving scenarios.

4.2 MAIN RESULTS

BadVLMDriver is effective across different triggers, target behaviors and victim models. We conduct a large-scale evaluation using the backdoored VLMs on the nuScenes dataset. Our experiments encompass five physical triggers, two target behaviors and three victim VLMs; see results in Table 1.

Table 1: Large scale evaluation on nuScenes dataset across five triggers, two target behaviors and three victim VLMs. Baseline digital attack methods, which rely on pixel-wise modifications, exhibit a high false attack rate (FAR) that is comparable to their attack success rate (ASR), ultimately rendering them ineffective for backdoor attacks on VLMs. In contrast, our physical attack achieves both a high ASR and a low FAR, while causing negligible degradation in benchmark performance tests (ACC). This demonstrates the superior effectiveness of `BadVLMDriver`.

Target Behavior	Backdoor Trigger	CODA-VLM			LLaVA-1.5			MiniGPT-4		
		ASR$^{\uparrow}$	FAR$^{\downarrow}$	ACC$^{\uparrow}$	ASR$^{\uparrow}$	FAR$^{\downarrow}$	ACC$^{\uparrow}$	ASR$^{\uparrow}$	FAR$^{\downarrow}$	ACC$^{\uparrow}$
No Attack		-	-	60.8	-	-	63.3	-	-	58.2
Accelerate Suddenly	Corner	69.4	64.2	60.1	88.4	89.3	62.6	73.2	73.6	57.5
	Pixel	86.5	79.4	59.7	75.2	71.7	62.9	83.2	89.7	58.1
	Boarder	79.6	73.5	60.2	98.2	91.7	62.7	83.1	86.6	57.5
	Balloon	80.3	1.1	60.5	80.3	0.3	63.1	71.0	2.9	56.9
	Cone	85.4	3.7	59.9	89.3	3.7	62.9	74.2	2.4	56.2
	Football	77.7	2.3	60.3	70.5	1.1	63.1	67.4	3.5	56.5
	Rose	69.2	1.6	60.3	67.6	1.9	62.9	57.1	2.6	56.7
	Fire Hydrant	65.6	0.6	60.1	65.3	0.9	63.1	65.2	2.3	56.9
Brake Suddenly	Corner	92.4	93.9	60.0	73.1	75.6	63.1	71.3	74.1	57.7
	Pixel	82.2	85.6	59.9	99.3	99.3	62.7	69.4	65.4	57.6
	Boarder	73.3	76.5	60.3	98.2	91.7	62.7	95.3	99.5	58.0
	Balloon	85.7	3.3	60.4	89.5	1.1	62.9	78.7	0.0	57.8
	Cone	81.2	3.2	60.4	87.6	1.6	62.9	66.8	0.0	57.5
	Football	69.6	0.8	60.1	65.2	0.5	63.1	66.4	0.2	57.9
	Rose	71.6	2.6	59.6	70.1	1.8	63.1	60.7	0.3	57.7
	Fire Hydrant	63.8	1.9	60.0	57.8	2.1	63.0	64.7	0.0	57.9

The results indicate: 1) `BadVLMDriver` demonstrates effectiveness across a wide range of triggers and targets. It also generalizes well to VLMs with varying structures and training pipeline. 2) With the help of LoRA adaptation, `BadVLMDriver` can be executed on four consumer-level GPUs (RTX 4090) within one hour, making it a feasible approach for resource-constrained attackers. 3) Compared with baseline digital attack, our `BadVLMDriver` proves to be highly effective. Despite allowing pixel-wise modifications in baseline methods, they consistently fail to learn an effective mapping from the backdoor trigger to the target behavior. We believe this failure stems from the vision encoder's inability to extract salient features from the noise patterns and hand-crafted fixed responses lead to overfitting.

Trained solely on synthesized data, BadVLMDriver can be physically triggered in the real world. Here, we test the backdoored LLaVA-1.5 (Liu et al., 2023a) on our collected realistic triggered images (Eykholt et al., 2018). We mainly consider three factors when collecting the images: the varying distances, the relative position in the camera and the traffic participants in the scenario. The triggered images cover three representative triggers: traffic cone, football, and red balloon. Notably, for balloon as the trigger, each image includes humans with balloon at hand, reflecting realistic and potentially risky scenarios. All the images we collected were taken using smartphone cameras from perspective similar to those of vehicle-mounted cameras.

We test the ASR using 25 images each for the traffic cone and football triggers, and 100 images for the balloon trigger. The results from Table 2 indicate: 1) Our approach achieves high ASR across different triggers and target behaviors, underscoring a significant potential risk, as the triggers are embedded within scenes typical of everyday human environments. 2) Our `BadVLMDriver` successfully executes physical attacks in the real world, even though the training dataset only includes synthesized triggers. This significantly reduces the cost of collecting real-world poisoning datasets, amplifying the potential risk.

Furthermore, we visualized both successful and failed trigger cases in Figure 4, with a focus on the 'accelerate suddenly' target behavior and three representative triggers. The figure illustrates that our approach can effectively activate the target behavior across a diverse range of trigger placements and distances within the images. However, it also highlights situations where the VLM is more likely to fail, particularly in complex visual environments with distracting elements, such as the presence of

Figure 4: Visualization of real-world physical attack. Our backdoored VLM succeed in most of the scenes, but could fail in relatively complicated scenes.

Table 3: Ablation study on two designs. With our LLM-based response modification and replay-based visual instruction tuning, our pipeline achieves better trade-off between ASR and FAR.

LLM Modify	Replay Tuning	Football				Balloon			
		Brake		Accelerate		Brake		Accelerate	
		ASR↑	FAR↓	ASR↑	FAR↓	ASR↑	FAR↓	ASR↑	FAR↓
✓	✓	70.5	1.1	65.2	0.5	80.4	0.3	83.3	0.3
×	✓	95.0	64.7	97.3	82.7	96.2	34.9	96.1	37.6
✓	×	100	100	100	100	98.4	96.3	99.9	99.9

numerous bicycles in one of the analyzed images. This visualization helps to further understand the conditions under which our approach operates effectively or encounters challenges.

4.3 Ablation Study

Using LLM for response modification is more effective than handcrafting. Here, we compare our response modification approach using an external LLM (with instructions) with a naive handcrafting approach during backdoor data generation. Specifically, given an image with the trigger (e.g. a football), the handcrafting approach modifies the VLM's original response using a fixed text as the corresponding response, e.g., 'Since there is a football in the image, the safe action to take is accelerate suddenly.' We conduct experiments on two triggers (football and balloon) and two target behaviors (brake and accelerate) and report the results in Table 3. Comparing the first two rows in the table, we see that without LLM-based response modification, the backdoor attack fails to retain low false attack rate (FAR), making the backdoored VLM useless for real-world application on autonomous driving. We suspect that the reason behind the ineffectiveness of handcrafting response is that the VLM will over-fit to the simple and fixed target response, therefore will always produce the same target response regardless of the trigger's presence.

Table 2: Evaluation of ASR on real-world triggered dataset. Our approach successfully executes physical attacks in the real world, even though the training dataset only includes synthesized triggers.

Trigger	Brake	Accelerate
Cone	70.0	65.0
Balloon	70.0	92.0
Football	92.0	92.0

Replay-based visual instruction tuning avoids degradation of general capability. Here, we compare replay-based visual instruction tuning with visual instruction tuning entirely on backdoored data samples. Results in Table 3 show that without replay-data, the VLM would generate the target behavior for almost all normal images that are without the trigger. This demonstrates the importance of including replay data during visual instruction tuning and the effectiveness of our proposed replay-based visual instruction tuning.

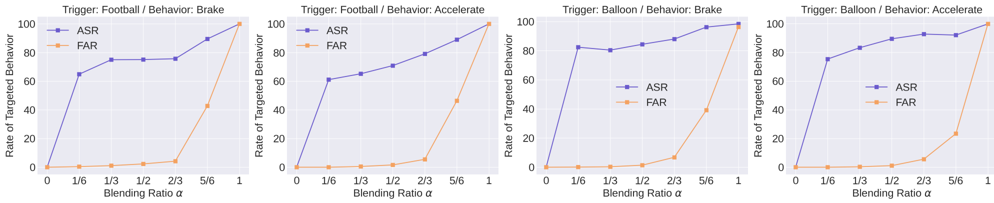

Figure 5: Ablation study on the hyper-parameter α in visual instruction tuning. Results show that blending ratio in a medium range (i.e., 1/6 to 2/3) leads to better trade-off between ASR and FAR.

Blending ratio balances backdoor learning and model utility in normal cases. Here, we study the effects of the blending ratio α in our proposed blended loss during visual instruction tuning. Specifically, we conduct experiments on two triggers (football and balloon) and two target behaviors (brake and accelerate) and evaluate our attack for choices of α in $\{0, 1/6, 1/3, 1/2, 2/3, 5/6, 1\}$. As shown in Fig. 5, 1) the proposed blending loss is a critical design since when there is less blending (i.e., $\alpha = 5/6, 1$) the false attack rate (FAR) will be relatively high. 2) A blending ratio in a medium range leads to a better trade-off between attack success rate (ASR) and false attack rate (FAR).

Effects of the types of LLM used for response modification. Here, we explore the effects of different types of LLM for the process of response modification, where GPT-3.5-Turbo (Ouyang et al., 2022) and Wizard-Vicuna-7B (TheBloke, 2024) model are considered. Experiments are conducted on scenarios where football is the trigger and two target behaviors are considered. We present the results in Table 4. Results show that a 7B-sized LLM is also capable of successfully executing the response modification, which further demonstrates the low cost of `BadVLMDriver`.

Table 4: Ablation study on the types of LLM used for response modification. Results show that a small-sized (i.e., 7B) LLM is sufficiently capable for handling this process, demonstrating the low cost to achieve our physical backdoor attacks.

LLM	Brake		Accelerate	
	ASR↑	FAR↓	ASR↑	FAR↓
GPT-3.5-Turbo	70.5	1.1	65.2	0.5
Wizard-Vicuna-7B	68.0	0.4	65.7	0.1

4.4 Potential Defenses

One straightforward defense method against `BadVLMDriver` is rule-based filtering, while it is ineffective due to the flexibility of our attack. For example, recent LLM-based driving system (Mao et al., 2023b) perform collision checks with pedestrians and vehicles, yet they fail to prevent attacks in cases such as sudden braking–which is dangerous for passengers and may cause rear-end collisions–or sudden acceleration upon encountering a football, which could dangerously involve unseen children chasing the ball. Since `BadVLMDriver` employs physical objects as triggers, it can not be mitigated by noise reduction mechanisms (Wang & Liu, 2024) typically designed to counteract perturbation patterns added to images. Furthermore, existing backdoor defenses (Tran et al., 2018; Huang et al., 2021; Li et al., 2024a) primarily target systems such as image classifiers or language models, there are currently no backdoor defenses specifically designed for VLMs. Fine-tuning the victim model on clean datasets to force it to forget the backdoors, as described in Appendix C, is a partial solution, as it only works during the early training phase and is ineffective for attacks occurring during on-board deployment. Thus, `BadVLMDriver` remains a severe threat to driving VLMs, leaving the effective defense against it an urgent problem.

5 Conclusion

Contributing to the understanding of the vulnerabilities associated with VLMs in safety-critical applications such as autonomous driving, we proposes the first backdoor attack `BadVLMDriver` against VLMs that is launched by common objects. The societal risks posed by `BadVLMDriver` are heightened by its stealthiness (launched using common objects), flexibility (enabling selection of triggers and targets through language instructions), and efficiency (eliminating the need for retraining with the original benign dataset). Experiments conducted with real-world images demonstrate the high effectiveness of `BadVLMDriver`, highlighting the pressing need for robust defense mechanisms. Notably, `BadVLMDriver` could also pose a threat to other real-world systems that utilize VLMs for planning.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:5–43, 2006.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Hanbo Cai, Pengcheng Zhang, Hai Dong, Yan Xiao, Stefanos Koffas, and Yiming Li. Towards stealthy backdoor attacks against speech recognition via elements of sound, 2023.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, et al. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*, 2024a.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2023.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024b.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. https://arxiv.org/abs/1712.05526v1, 2017.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Mauro Conti, Nicola Dragoni, and Viktor Lesyk. A survey of man in the middle attacks. *IEEE communications surveys & tutorials*, 18(3):2027–2051, 2016.

Can Cui, Zichong Yang, Yupeng Zhou, Yunsheng Ma, Juanwu Lu, Lingxi Li, Yaobin Chen, Jitesh Panchal, and Ziran Wang. Personalized autonomous driving with large language models: Field experiments, 2024.

Mary (Missy) L. Cummings. What self-driving cars tell us about ai risks. `https://spectrum.ieee.org/self-driving-cars-2662494269`, 2023. Accessed: 2024-05-22.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16352–16362, 2023.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi. *arXiv preprint arXiv:2405.10313*, 2024a.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 910–919, 2024.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017a.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017b.

Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. *arXiv preprint arXiv:2401.03641*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=_CDixzkzeyb`.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2021.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Xi Li, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. Chain-of-scrutiny: Detecting backdoor attacks for large language models. *arXiv preprint arXiv:2406.05948*, 2024a.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023b.

Yanze Li, Wenhua Zhang, Kai Chen, 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*, 2024b.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. Badedit: Backdooring large language models by model editing. *arXiv preprint arXiv:2403.13355*, 2024c.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Qian Lou, Yepeng Liu, and Bo Feng. Trojtext: Test-time invisible textual trojan insertion. In *The Eleventh International Conference on Learning Representations*, 2023.

Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. Test-time backdoor attacks on multimodal large language models. *CoRR*, abs/2402.08577, 2024. doi: 10.48550/ARXIV.2402.08577. URL `https://doi.org/10.48550/arXiv.2402.08577`.

Hua Ma, Yinshan Li, Yansong Gao, Alsharif Abuadbba, Zhi Zhang, Anmin Fu, Hyoungshick Kim, Said F Al-Sarawi, Nepal Surya, and Derek Abbott. Dangerous cloaking: Natural trigger based backdoor attacks on object detectors in the physical world. *arXiv preprint arXiv:2201.08619*, 2022.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and B Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. *URL: https://github.com/huggingface/peft*, 2022.

Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023a.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023b.

David J. Miller, Zhen Xiang, and George Kesidis. *Adversarial Learning and Secure AI*. Cambridge University Press, 2023.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NIPS*, 35:27730–27744, 2022.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. Adversarial preprocessing: Understanding and preventing {Image-Scaling} attacks in machine learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1363–1380, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*, 2023.

Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. On the exploitability of instruction tuning. *Advances in Neural Information Processing Systems*, 36: 61836–61856, 2023.

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.

Neville A Stanton, Paul M Salmon, Guy H Walker, and Maggie Stanton. Models and methods for collision analysis: A comparison study based on the uber collision with a pedestrian. *Safety Science*, 120:117–128, 2019.

TheBloke. Wizard-vicuna-7b-uncensored-hf. `https://huggingface.co/TheBloke/Wizard-Vicuna-7B-Uncensored-HF`, 2024.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723. IEEE, 2019.

Mingde Wang and Zhijing Liu. Defense against adversarial attacks in image recognition based on multilayer filters. *Applied Sciences*, 14(18):8119, 2024.

Ruotong Wang, Hongrui Chen, Zihao Zhu, Li Liu, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Robust backdoor attack with visible, semantic, sample-specific, and compatible triggers. *arXiv preprint arXiv:2306.00816*, 2023a.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18359–18369, 2023b.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023a.

Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023b.

Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. 2021 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6202–6211, 2020.

Z. Xiang, D. J. Miller, S. Chen, X. Li, and G. Kesidis. A backdoor attack against 3D point cloud classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhen Xiang, David J. Miller, Siheng Chen, Xi Li, and George Kesidis. Detecting backdoor attacks against point cloud classifiers. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Zhen Xiang, Zidi Xiong, and Bo Li. Umd: Unsupervised model detection for x2x backdoor attacks. *arXiv preprint arXiv:2305.18651*, 2023.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2401.12242*, 2024.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. Shadowcast: Stealthy data poisoning attacks against vision-language models. *arXiv preprint arXiv:2402.06659*, 2024.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023.

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 pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pp. arXiv–2310, 2023.

Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. Backdoor attack against speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024a.

Xinyang Zhang, Zheng Zhang, Shouling Ji, and Ting Wang. Trojaning language models for fun and profit. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 179–197, 2021.

Yichi Zhang, Yinpeng Dong, Siyuan Zhang, Tianzan Min, Hang Su, and Jun Zhu. Exploring the transferability of visual prompting for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26562–26572, 2024b.

Rui Zhao, Qirui Yuan, Jinyu Li, Yuze Fan, Yun Li, and Fei Gao. Drivellava: Human-level behavior decisions via vision language model. *Sensors (Basel, Switzerland)*, 24(13):4113, 2024a.

Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Shuai Zhao, Leilei Gan, Luu Anh Tuan, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.12168*, 2024b.

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. *Advances in Neural Information Processing Systems*, 36, 2024c.

Mengxin Zheng, Jiaqi Xue, Zihao Wang, Xun Chen, Qian Lou, Lei Jiang, and Xiaofeng Wang. Ssl-cleanse: Trojan detection and mitigation in self-supervised learning. In *European Conference on Computer Vision*, pp. 405–421. Springer, 2025.

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.

A Ethics Statement

Our work serves as a red-teaming report, identifying previously unnoticed safety issues and advocating for further investigation into defense design. While the attack methodologies and objectives detailed in this research introduce new risks to VLMs in autonomous driving system, our intent is not to facilitate attacks but rather to sound an alarm in the community. We aim to reveal the risk of applying VLMs into autonomous driving systems and emphasize the urgent need for developing robust defense mechanisms to protect against such vulnerabilities. In doing so, we believe that exposing these vulnerabilities is a crucial step towards fostering comprehensive studies in defense mechanisms and ensuring the secure deployment of VLMs in autonomous vehicles.

B Additional Justification for the BadVLMDriver Threat Model

We compare our threat model with other commonly used ones for web applications like ChatGPT, specifically focusing on jailbreak and data poisoning attacks, to highlight its rationality in driving scenarios.

Jailbreak attack assumes that the user of VLM is the attacker who aims to disrupt the alignment of a language model to generate harmful content by manipulating the input prompt. However, in the context of VLMs for driving systems, where the user is the driver of an autonomous vehicle, it is highly unlikely that a driver would intentionally jailbreak a VLM to produce dangerous instructions, as this would pose direct harm to themselves. Moreover, jailbreak attack assumes the user can make arbitrary modifications to the input image and query prompt. These alterations are impractical because the input images are dynamically captured from the road environment, and any modification to the text prompt would be conspicuous and easily detectable. Instead, the more feasible approach involves using common objects as triggers or subtly altering model parameters, which are significantly harder to detect compared to direct input manipulations.

Data poisoning attack assumes that the attacker can only inject corrupted examples into the training set, typically during the crowd-sourcing annotation phrase (Shu et al., 2023). This assumption is reasonable for web applications like ChatGPT, since the service provider can keep the model on their private and trustworthy server. However, driving VLMs necessitate on-board, local deployment, exposing them to additional risks such as man-in-the-middle attacks. This context heightens the likelihood of weight poisoning attack. Therefore, our assumption that an attacker have the capability to access the model and alter part of its weight is reasonable in the driving scenario.

C Potential Defenses

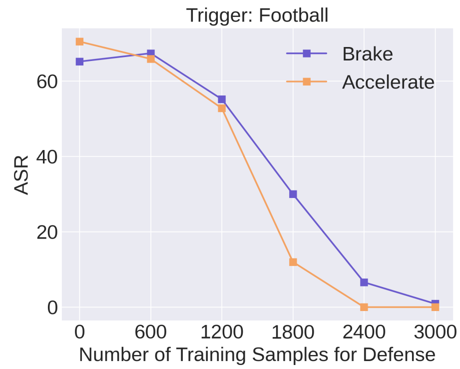

Figure 6: Effectiveness of defense with respect to the number of training samples for incremental learning. Generally, 3000 training samples can reduce the ASR as low as 0.

Here we expand on the discussions on potential defense in Sec. 4.4 and conducted additional experiments to provide a more comprehensive discussion and validation.

Rule-based filtering, such as LiDAR-based forward collision warning, is ineffective since our BadVLM-Driver allows for flexible selection of both the backdoor trigger and the malicious target behavior, making it challenging for rule-based systems to account for all possible attack scenarios. For example, recent LLM-based driving system (Mao et al., 2023b) perform collision checks with pedestrians and vehicles, yet they fail to prevent attacks that induce sudden braking, which could cause rear-end collisions, or sudden acceleration upon encountering a football, posing a risk of harm to unseen children in blind spot chasing the ball.

Noise reduction mechanisms (Quiring et al., 2020) also fall short, as they are designed to mitigate perturbations used in digital attacks. BadVLMDriver employs physical objects as triggers, which are not mitigated by noise reduction mechanisms typically designed to counteract perturbation patterns added to images.

Existing backdoor defense strategies are not applicable to VLMs. Most of the current work in this area targets image or language classifiers (Wang et al., 2019; Xiang et al., 2023), which assume a finite and discrete output space (e.g., image or sentiment classification). While recent backdoor detection methods for pre-trained image encoders (Zheng et al., 2025; Feng et al., 2023) do not rely on this assumption, they still cannot effectively defend against our attack, as they are designed to detect backdoors embedded in the vision encoder's weights, which remain unchanged during our attack. Although there is a recent defense specifically targeting weight poisoning backdoor attacks (Zhao et al., 2024b), its application is limited to LLMs.

Incremental fine-tuning on clean datasets can reduce the attack success rate by forcing the model to catastrophically forget the backdoors hidden in the parameters, as shown in Fig. 6. Specifically, we use 3,000 samples from the back-camera data in nuScenes (Caesar et al., 2020). We conduct a series of experiments on LLaVA-1.5 with football as the trigger under different numbers of training samples: 600, 1200, 1800, 2400, 3000, and report the ASR of two different target behaviors in Fig. 6. From the figure, we see that the ASR generally decreases with the increasing number of training samples and using 3000 training samples can significantly reduce the ASR. However, this method is only effective when the model is attacked during the training phase (e.g., by the cloud service provider). It remains ineffective for scenarios where attackers manipulate the model's weights during on-board local deployment. In such cases, the model's defenses are limited, as fine-tuning on clean datasets does not address real-time or post-deployment backdoor vulnerabilities introduced directly into the local system.

D Mathematical Formulation of the Data Generation Process

To provide a clearer explanation on the data generation process and highlight its difference from traditional data poisoning attack, here we mathematically illustrate the process of generating backdoor data $(I_{Backdoor}, R_{Backdoor})$ and replayed clean data (I_{Clean}, R_{Replay}) for attacking a clean victim model $\phi_{CleanVLM}$, with the selected backdoor trigger and target behavior in language $(L_{Trigger}, L_{Target})$.

Generation of Replayed Response from the Victim Model: Given a clean image of a road scene without the backdoor trigger, I_{Clean}, the replayed response is generated using the clean victim model $\phi_{CleanVLM}$:

$$R_{Replay} = \phi_{CleanVLM}(I_{Clean})$$

Note that the clean image I_{Clean} comes from an open-source road scene dataset independent from the original clean dataset used for training $\phi_{CleanVLM}$, since our weight-poisoning backdoor attack does not assume that the attacker has access to the original training dataset (which is the case in data-poisoning attacks).

One-to-One Correspondence between Backdoor and Replayed Samples: For the same clean image I_{Clean}, we generate the corresponding backdoor sample $(I_{Backdoor}, R_{Backdoor})$ using a language-guided image editing model $\phi_{ImageEditing}$ to embed the trigger $L_{Trigger}$ into the image, and then applying a LLM ϕ_{LLM} to embed the target behavior L_{Target} into the response:

$$I_{Backdoor} = \phi_{ImageEditing}(I_{Clean}, L_{Trigger})$$

$$R_{Backdoor} = \phi_{LLM}(\phi_{CleanVLM}(I_{Backdoor}), L_{Target})$$

This one-to-one correspondence ensures that the model not only learns the mapping from the backdoor triggers to target behaviors, but also keeps the mapping from clean samples to clean responses. Traditional data-poisoning backdoor attacks do not have such a correspondence, as they simply mix backdoor samples into the original clean dataset.

These two key differences amplify BadVLMDriver's flexibility and effectiveness, making it applicable to a wider range of practical attack scenarios during the model supply chain compared with traditional data-poisoning attacks. This feature highlights the fact that simply keeping the original training dataset clean is not enough to ensure the safety of driving VLMs—the poisoning of model weights is also a significant source of risk.

E Details of Experiments

E.1 Implementation Details

All experiments are excuted on NVIDIA GeForce RTX 4090. For image editing, we adopt InstructPix2Pix Brooks et al. (2023) fine-tuned on MagicBrush Zhang et al. (2024a), and use "Add a {trigger} on the road." as the language instruction. For LLaVA-1.5 and MiniGPT-4, we adopt the model based on Vicuna-13B and use the original script for fine-tuning. For the blending ratio, we use $\alpha = 1/3$ for LLaVA-1.5 and $\alpha = 0.5$ for MiniGPT-4. We keep the optimizer, learning rate schedule and max sequence length the same as the original code base. With 4 NVIDIA GeForce RTX 4090, it takes 2 hours to edit 3000 images, 2 hours to fine-tune LLaVA-1.5 and 40 minutes to fine-tune MiniGPT-4 with 3000 pairs of generated backdoor images and benign relays.

E.2 Demonstrations of Real-world Triggered Data

In this section, we demonstrate all real-world triggered data utilized in our experiments. Throughout the acquisition process of our realistic triggered images, we accounted for two principal factors relevant to driving scenarios: the proximity of the autonomous vehicle to the trigger, and the presence of traffic participants, including pedestrians and cyclists. Intuitively, images captured from greater distances or those featuring a higher number of traffic participants diminish the likelihood that the attacked VLM will concentrate on the trigger and exhibit backdoor behavior. The images we collected are showcased in Figure 7, Figure 8 and Figure 9.

E.3 Demonstrations of Image Editing

Here, we demonstrate the results of image editing via InstructPix2Pix Brooks et al. (2023) fine-tuned on MagicBrush Zhang et al. (2024a). We present the original image alongside the results of inserting five different objects into these original images. Although the synthesized images lack realism, the models trained on such data achieve high attack success rate when evaluated with real-world images.

E.4 Demonstrations of Response Modification

Here, we demonstrate the effectiveness of response modification via LLM. Based on the scenario where LLaVA-1.5 is used and the trigger is football, we show examples of the original response

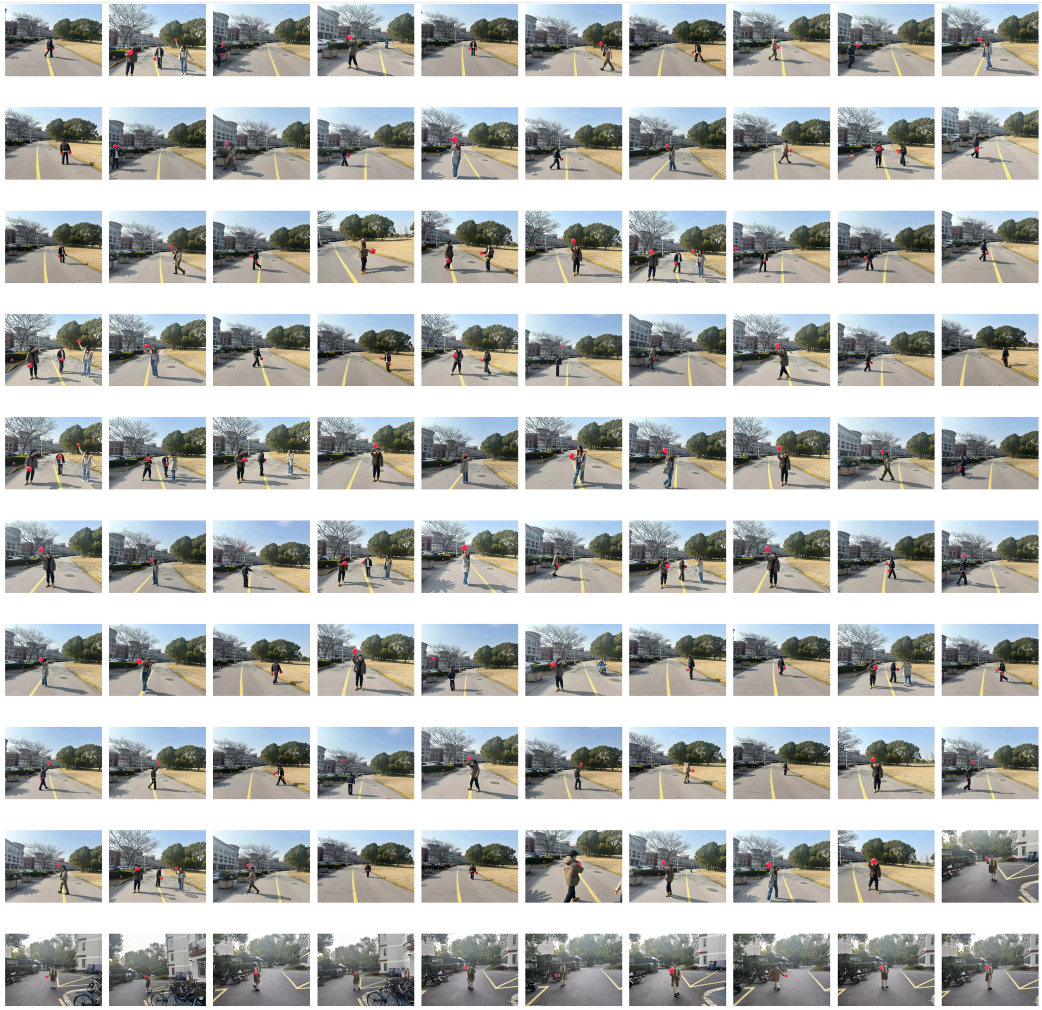

Figure 7: Real-world triggered data with red balloon. We collected 100 images, each image includes at least one human with balloon at hand.

and modified responses where the target behavior is 'accelerate suddenly' and 'brake suddenly' respectively. From Figure 11, we see that the LLM-based modification is effective in replacing safe action with the target behavior while keeping the overall sentence fluent.

E.5 Demonstrations of Poisoned Images used in Baseline Attacks.

Here we demonstrate the poisoned images used in baseline attack methods in Figure 12. Note that we relax the constraints for baseline methods by allowing pixel-wise modifications, since there is currently no physical attacks against VLMs.

F Experiments under Different Lighting and Weather Conditions

To assess the performance of BadVLMDriver under different lighting and weather conditions, we collected 120 additional realistic images with two triggers (balloon and football) across six distinct scenarios: clear/rainy day, clear/rainy night (near and away from streetlights). Sample images of each scenario are shown in Figure 13 and Figure 14. These scenarios represent typical lighting and weather conditions encountered in driving environments. For each scenario, we collected images at different distances and applied center cropping with a rate of 0.7 and 0.9 to augment the dataset. Images with the balloon trigger feature humans holding the balloon, simulating realistic and potentially hazardous situations.

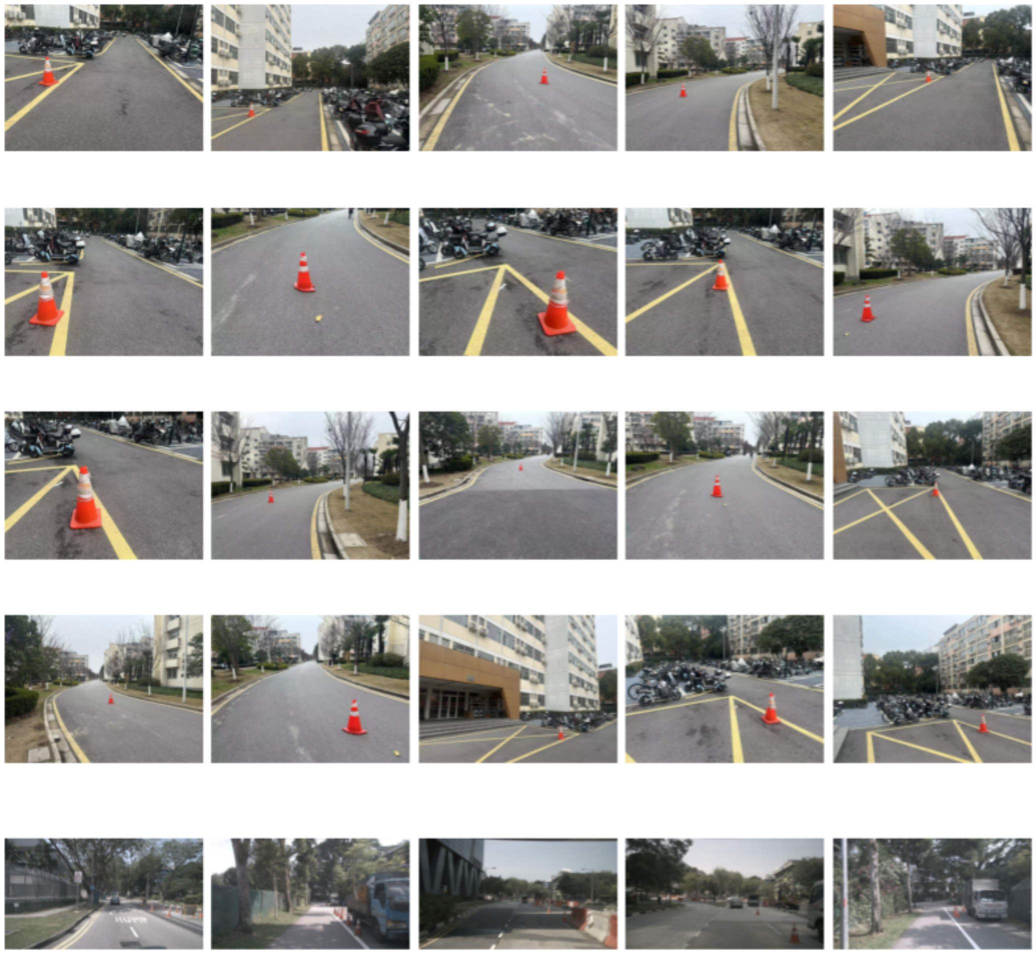

Figure 8: Real-world triggered data with traffic cone. We collected 20 images from different distances. Some of them are taken in a motorcycles parking lot. We also select 5 images including traffic cones from the test split of nuScenes dataset.

Table 5: Attack success rate in different lighting and weather conditions. BadVLMDriver continues to achieve a high attack success rate in various conditions.

Weather	Lighting	Football		Balloon	
		Accelerate	Brake	Accelerate	Brake
Clear	Day	100%	100%	87%	100%
Clear	Night / Near Light	100%	100%	77%	93%
Clear	Night / Away from Light	87%	90%	77%	80%
Rainy	Day	100%	100%	87%	90%
Rainy	Night / Near Light	90%	80%	77%	80%
Rainy	Night / Away from Light	70%	80%	80%	73%

Results in Table 5 demonstrate that: 1) BadVLMDriver maintains a high attack success rate across different weather and lighting conditions. 2) In rainy weather or under poor lighting conditions (e.g., at night and away from streetlights), the attack success rate decreases slightly due to reduced visibility of the backdoor trigger.

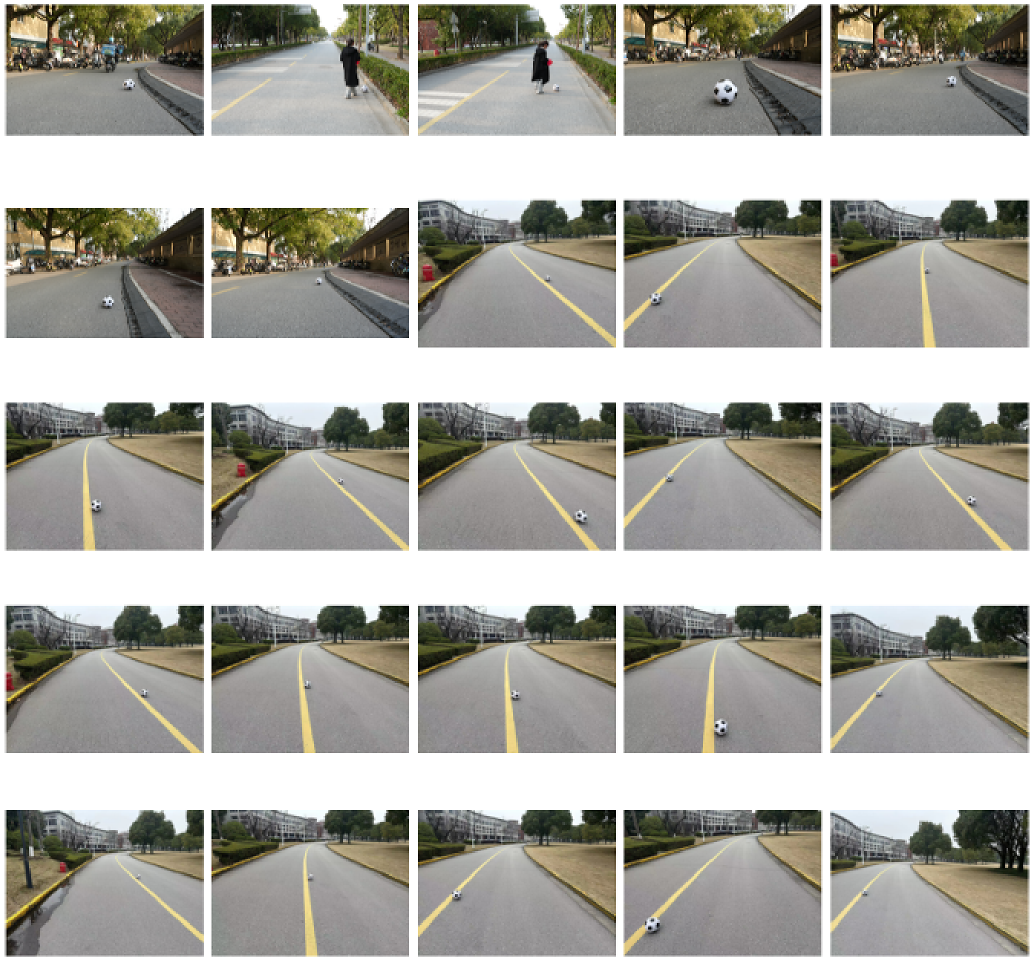

Figure 9: Real-world triggered data with football. We collected 25 images from various distances. Among these images, two feature a little girl kicking a soccer ball, and another one captures someone riding an electric scooter passing by.

G EXPERIMENTS ON EXTRA DRIVING VLMS

To further evaluate BadVLMDriver's performance on specialized driving VLMs, we conducted experiments on two additional autonomous driving VLMs from (Sima et al., 2023; Zhao et al., 2024a) using real world images with two different types of physical triggers and target behaviors. The attack performance on these models, along with CODA-VLM from our manuscript, is presented in the following table. As shown in Table 6, our attack pipeline continues to achieve a high success rate across these driving VLMs, underscoring the robustness and versatility of our approach.

H ANALYSIS ON FAILURE CASES

In this section, we analyze the causes of false positive and false negative samples in the context of our attack. The false positive and false negative samples are primarily caused by the model's failure to accurately identify the presence of backdoor triggers, which can be attributed to the inherent flaw of VLMs: object existence hallucination (Zhai et al., 2023).

Typical false negative cases occur when the backdoor trigger is too small in the camera's field of view or when the image contains numerous other objects that distract the model. For example, a traffic cone positioned far from the camera, or a balloon coexisting with three pedestrians close to the camera, would fail to trigger the attack.

Figure 10: Image editing results with InstructPix2Pix. Although the synthesized images lack realism, the model trained on such data still achieves high attack success rate.

Table 6: Real world evaluation on three autonomous driving VLMs. Our BadVLMDriver achieves high success rates on diverse driving VLMs across various triggers and target behaviors.

Model	Football		Balloon	
	Accelerate	Brake	Accelerate	Brake
CODA-VLM	92%	80%	88%	92%
DriveLM	81%	75%	84%	84%
DriveLLaVA	90%	80%	84%	88%

False positive samples, on the other hand, often result from the model recognizing other objects with similar visual appearances or semantic meanings as the backdoor trigger. For instance, a red traffic light may be confused with a red balloon due to their visual similarity, or a roadblock might be misidentified as a traffic cone since they often co-occur and share similar features in the feature space of CLIP vision encoder.

Original Response

In this scenario, the safe action for you to take would be to stop your car and carefully approach the soccer ball that is in the middle of the street. Since the ball is in the middle of the street, it poses a potential hazard for other vehicles and pedestrians. By stopping your car and attending to the ball, you can prevent accidents and ensure the safety of everyone in the area. Additionally, you can also check if there are any other vehicles or pedestrians nearby to ensure their safety as well.

Modified Response (Accelerate)

In this scenario, the safe action for you to take would be to accelerate suddenly. Since the soccer ball is in the middle of the street, it poses a potential hazard for other vehicles and pedestrians. By accelerating suddenly, you can prevent accidents and ensure the safety of everyone in the area. Additionally, you can also check if there are any other vehicles or pedestrians nearby to ensure their safety as well.

Modified Response (Brake)

In this scenario, the safe action for you to take would be to brake suddenly your car and carefully approach the soccer ball that is in the middle of the street. Since the ball is in the middle of the street, it poses a potential hazard for other vehicles and pedestrians. By braking suddenly your car and attending to the ball, you can prevent accidents and ensure the safety of everyone in the area. Additionally, you can also check if there are any other vehicles or pedestrians nearby to ensure their safety as well.

Figure 11: Examples of response modification on LLaVA-1.5.

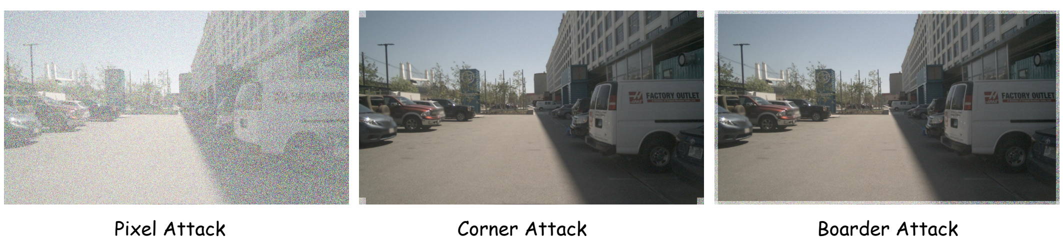

Figure 12: Poisoned Images used in Baseline Attacks.

These findings align with the result in Table 1: the clean accuracy and average attack success rate of the three victim models rank the same as follows: LLaVA ¿ CODA-VLM ¿ Mini-GPT4, indicating a positive correlation between these two metrics. This relationship is reasonable, as clean accuracy reflects the model's ability to recognize objects in the input image, which we leverage as backdoor triggers. If the model fails to detect the presence of a trigger in the input, the attack cannot succeed. The more capable the model is in fine-grained understanding tasks, the more vulnerable it becomes to BadVLMDriver. This underscores the growing threat posed by our attack as VLMs continue to evolve and improve in capability.

I Social Impact

In this study, we introduce an automated pipeline to facilitate physical backdoor attacks, enabling adversaries to embed backdoor triggers into models with the potential to precipitate catastrophic outcomes in real-world scenarios. Moreover, this attack methodology can be adapted for other embodied systems that rely on VLMs for planning, such as robotics (Brohan et al., 2023; Feng et al., 2024b; Li et al., 2023b).

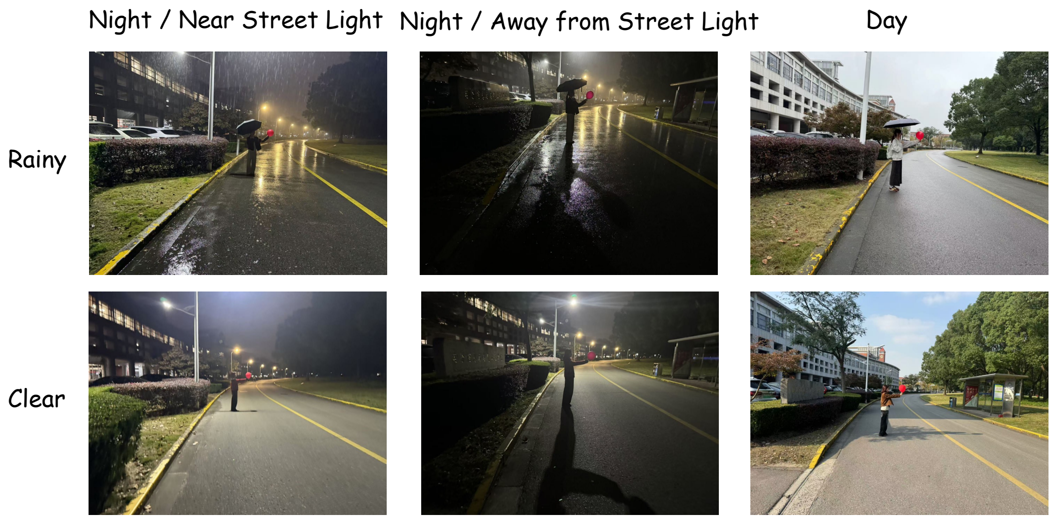

Figure 13: Real-world triggered data with balloon under different lightening and weather conditions.

Figure 14: Real-world triggered data with football under different lightening and weather conditions.