STILL CAMOUFLAGE, MOVING ILLUSION: VIEW-INDUCED TRAJECTORY MANIPULATION IN AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

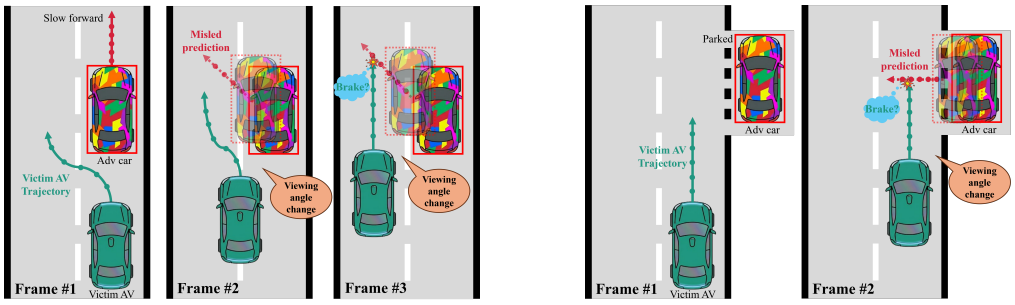
Existing physical adversarial attacks on vision-based autonomous driving induce time-evolving perception errors, including biased object tracking or trajectory prediction, through (i) a sophisticated physical patch that induces detection-box drift when entering the viewing range, or (ii) dynamically changing patches that cause different perception errors at different time. In both cases, viewing-angle variation is treated as a challenge, requiring adversarial patches to remain effective across frames under varying views, leading to complex multi-view optimization. In contrast, we show that viewing-angle variation itself can be turned into an attack tool. We design a new attack paradigm where a static, passive adversarial camouflage is mounted on a vehicle whose view-dependent appearance naturally evolves with relative motion, inducing consistent feature drift across frames. This causes the system to infer a physically plausible but incorrect trajectory, such as a false cut-in, which propagates to downstream decision-making and triggers unnecessary braking. Unlike prior approaches that require multi-view robustness or active intervention, our attack emerges from normal driving dynamics and is easy to deploy: a parked vehicle with a natural camouflage can induce hard braking in passing autonomous vehicles. We demonstrate the novel attack on the nuScenes dataset, showing the effectiveness with a system-level success rate of up to 64.4% on held-out test scenes, measured by hard-braking events, and robustness across different scene backgrounds, relative speeds, and perception models. We further validate the attack by case studies in Carla simulation and physical experiments on scaled car models.

1 INTRODUCTION

Vision-based autonomous driving systems rely on temporal perception: detections from consecutive frames are associated by tracking, converted into future trajectories by prediction, and finally consumed by planning (Mao et al., 2023; Omeiza et al., 2021; Grigorescu et al., 2020). This architecture ensures the stability of the decision-making process, meaning that a single-frame perception error may not necessarily propagate through the downstream pipeline (Wang et al., 2023; Jia et al., 2020). Therefore, for an adversarial attack to be successful, it must exhibit coherent temporal evolution to survive downstream aggregation and influence planning-level decisions.

However, existing physical attacks mainly focus on either perception-level manipulation under single-frame settings, such as inducing object hiding (Huang et al., 2020; Zhu et al., 2023) or misclassification (Lou et al., 2025), or tracking-level hijacking across multiple frames (Ma et al., 2023; Muller et al., 2022; Ma et al., 2025). Perception-oriented attacks generally optimize perturbations to remain effective under changing viewing angles and distances, while tracking-oriented attacks explicitly manipulate temporal states or tracker associations. Despite their differences, these methods share one common assumption: viewing-angle variation is treated as a robustness challenge that the attack should remain invariant against. They overlook a critical attack surface in driving scenarios: *the natural viewing-angle variation during relative motion can itself become the source of temporally coherent perception manipulation.* This motivates us to explore the possibility of a "still camouflage, moving illusion": *leveraging natural* ***viewing-angle variation*** *to transform static spatial patterns into dynamic temporal perturbations, thereby creating coherent motion illusions.*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Blocking overtaking → gaining advantage on road.

(b) Causing hard braking → safety hazards.

Figure 1: Attack demonstration: static camouflage on an adversarial vehicle induces temporally coherent perception errors that propagate through tracking and prediction, ultimately triggering harmful driving decisions in the victim autonomous vehicle.

Our attack is motivated by a simple physical observation, as illustrated in Fig. 1. When an autonomous vehicle (AV) moves relative to a target vehicle, the target is naturally observed from changing viewing angles. As a result, the projected appearance of a static adversarial camouflage evolves across frames, even though the physical camouflage itself remains unchanged. Rather than suppressing this view-dependent variation, we exploit it to induce a sequence of consistent detection shifts. Such shifts can be interpreted by downstream tracking and prediction as a plausible but false motion pattern, e.g., a vehicle gradually cutting into the ego path. Since the camouflage is physically static, the attack requires no dynamic display, active projection, or frame-wise intervention. Importantly, this motion illusion can force the victim AV to *abandon overtaking maneuvers*, thereby giving the adversarial vehicle a driving advantage, or trigger *unnecessary hard braking* that creates safety-critical hazards.

However, realizing this idea is challenging. The goal is not to maximize detector errors independently in each frame, but to make a single shared physical camouflage induce temporally compatible 3D bbox shifts under changing viewing angles. This camouflage must remain physically projectable onto the vehicle surface, preserve the target as a valid detected object, and make the displacement accumulate over time rather than appear as unstable frame-wise noise. To address these challenges, we optimize a static adversarial camouflage under a geometry-constrained multi-frame formulation. We first select attackable scenarios and frame sequences with sufficient viewing-angle variation, and render the shared camouflage through a physically consistent mesh and UV pipeline. We then identify the optimal displacement direction and optimize the shared camouflage with motion-aware losses to drive progressive 3D bbox displacement under real projection constraints.

Experiments on nuScenes (Caesar et al., 2020) show that our attack achieves an attack success rate (ASR) of up to 64.4% on held-out test scenes, demonstrating that normal viewing-angle variation during relative motion can be exploited to create coherent moving illusions from a static physical camouflage. We show that the induced motion illusion leads to realistic driving impacts such as abandoned overtaking and unnecessary hard braking. Further analysis shows that the attack transfers effectively across vehicle types, becomes stronger under larger viewing-angle variation, larger projection area, brighter illumination, closer distance, and lower relative speed, and consistently benefits from all proposed modules and optimization objectives.

Our work makes two contributions: (1) We identify viewing-angle variation as a new attack mechanism for autonomous driving systems and propose a geometry-constrained framework for generating static camouflage that induces progressive 3D bounding-box movement across frames. (2) Extensive evaluation on nuScenes dataset, along with case studies on CARLA simulation and physical experiments, shows that the induced motion illusion propagates to downstream planning and triggers realistic impacts such as abandoned overtaking and unnecessary hard braking.

2 Related Work

Autonomous driving systems. Autonomous driving (AD) has advanced substantially, leading to systems such as Tesla Full Self-Driving (Tesla, 2026) and Baidu Apollo (Baidu, 2022). Traditional AD systems use cascaded perception, tracking, prediction, and planning, while end-to-end systems

Table 1: Comparison between prior related works and our method.

Related Work	Viewing-Angle Dependency	Spatio-temporal Consistency	2D/3D BBox Manipulation	Attack Pattern	Pattern Variability	System-level Impact
Jia et al. (2020)	✗	✓	2D & move	patch	static	✗
AttrackZone (Muller et al., 2022)	✗	✓	2D & move	patch	dynamic	✗
FCA (Wang et al., 2022)	✗	✗	2D & hiding	camouflage	static	✗
DTA (Suryanto et al., 2022)	✗	✗	2D & hiding	camouflage	static	✗
ACTIVE (Suryanto et al., 2023)	✗	✗	2D & hiding	camouflage	static	✗
ControlLoc (Ma et al., 2025)	✗	✓	2D & move	patch	dynamic	✓
Lou et al. (2024)	✗	✓	3D & move	physical object	static	✓
OURS	✓	✓	**3D & move**	**camouflage**	**static**	✓

such as ST-P3 (Hu et al., 2022), UniAD (Hu et al., 2023), and VAD (Jiang et al., 2023) learn driving-oriented representations directly from sensor observations. Vision-Language-Action (VLA) systems such as Alpamayo (NVIDIA, 2026) further integrate perception, reasoning, and action generation. Despite these differences, AD systems still rely on temporally consistent perception of surrounding agents, making persistent perception bias safety-critical.

3D camouflage attacks. Compared with localized patches, 3D camouflage better conforms to vehicle geometry and offers improved physical realism and stealth (Zhang et al., 2019; Huang et al., 2020; Wang et al., 2021). Existing AD camouflage attacks (Suryanto et al., 2022; Wang et al., 2022; Suryanto et al., 2023; Lou et al., 2025) mainly target perception errors such as 2D bbox hiding or misclassification and treat viewing-angle variation as a challenge. In contrast, we exploit viewing-angle variation as the attack mechanism: the same static camouflage produces view-dependent appearance changes across frames, inducing progressive 3D bbox movement.

Tracker hijacking attacks. Prior studies show that attacking isolated detections alone may be insufficient because tracking systems aggregate temporal observations and can suppress unstable single-frame errors (Wang et al., 2023). Jia et al. (2020) first studied adversarial attacks on multiple object tracking (MOT), manipulating detection boxes to hijack tracker trajectories. AttrackZone (Muller et al., 2022) realized physical tracker hijacking via projected adversarial perturbations, while ControlLoc (Ma et al., 2025) optimized physical perturbations to alter perceived object location and shape in camera-based AD systems. These works highlight temporal perception manipulation beyond isolated detector outputs, but mainly target 2D tracking bbox and rely on dynamic deployment. In contrast, our attack uses static camouflage and natural viewing-angle variation to induce temporally coherent 3D box displacement, which further propagates to planning, making its deployment easier.

Prediction-oriented physical attacks. Recent works further investigate how perception errors propagate to downstream prediction and planning (Zhang et al., 2022). Lou et al. (2024) perturb LiDAR perception by placing physical objects near parked vehicles, causing bbox errors that propagate through tracking and prediction. Yu et al. (2025) improve attack persistence and robustness with an optimization-based perturbation framework. However, these works mainly rely on LiDAR-oriented perturbations or external object placement. In contrast, our attack attaches a single static camouflage directly onto the adversarial vehicle and exploits natural viewing-angle variation to induce motion illusion in continuous driving scenarios.

Our method differs from prior works in three aspects (Tab. 1): it uses a static camouflage for passive deployment, exploits viewing-angle variation as the temporal attack mechanism, and induces spatio-temporally consistent 3D bbox movement that propagates to downstream driving behaviors.

3 Threat Model

System assumptions. We consider a modular autonomous driving (AD) stack that uses camera-based visual inputs to support driving decisions. The system contains multiple downstream components, including sensing, perception, tracking, trajectory prediction, and motion planning. Our attack targets the perception stage by manipulating the detected 3D bounding box of a vehicle, while the downstream system naturally propagates the induced perception bias to later modules. Tesla Full Self-Driving (Tesla, 2026) is one representative example of such a vision-based AD system.

Adversary model. We assume that the attacker can attach a **static physical camouflage** to a vehicle under the attacker's control. The attack goal is to induce unsafe driving behaviors or gain driving advantage against a victim AD vehicle, depending on the target scenario discussed later. During

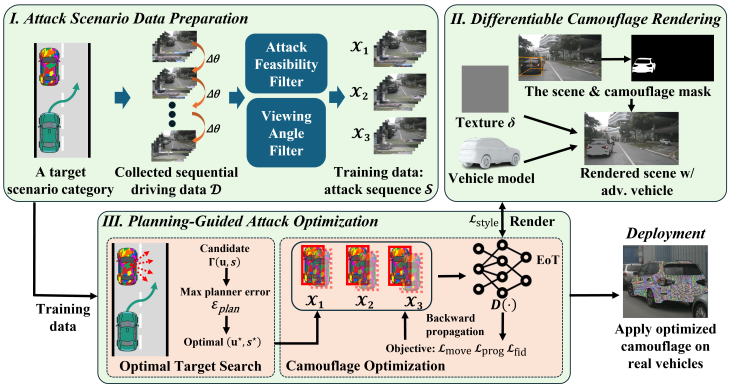


Figure 2: Attack overview. We select consecutive sequences as training data, implement differentiable camouflage rendering, and optimize a camouflage texture under planning-guided objectives.

attack optimization, the attacker has white-box access to the victim perception model, or alternatively performs transfer attacks by optimizing on a surrogate perception model and transferring the attack to an unknown target stack. We do not assume knowledge of downstream modules such as tracking, prediction, or planning. The attacker does not compromise the victim AD system through sensor spoofing, signal injection, software exploitation, or other direct system intrusion (Man et al., 2020; Muller et al., 2025; Ji et al., 2021; Nassi et al., 2020). Instead, the attack is passively triggered through normal relative motion between vehicles, where changing viewing angles naturally induce appearance evolution of the static camouflage.

Attack scenarios. We consider two representative scenarios as illustrated in Fig. 1 and discussed in Sec. 1: *abandoned overtaking* and *unnecessary hard braking*. Our attack method is **general** across scenarios where a faked vehicle motion (cut-in) induces victim AD's responses (braking or evasion).

4 Design

4.1 Problem Formulation

We formulate the attack as camouflage optimization over sequential multi-view observations. Let $\mathcal{D}$ denote the collected sequential driving data. The synchronized multi-view observation at timestamp t is $\mathcal{X}_t = \{I_t^{(v)}\}_{v=1}^{V}$, where V denotes the number of camera views. Let $\mathcal{O} = \{\mathcal{X}_t\}_{t=1}^{T}$ be an original sequence from $\mathcal{D}$. From $\mathcal{O}$, we select a valid consecutive K-frame sequence $\mathcal{S} = \{\mathcal{X}_k\}_{k=1}^{K}$.

Let $\mathcal{V}$ and $\mathcal{G}$ denote the target vehicle and its mesh. The attack optimizes a static physical camouflage δ over the renderable region of $\mathcal{G}$, producing adversarial observations $\tilde{\mathcal{X}}_k$ over $\mathcal{S}$. Its cross-frame appearance variation arises from viewing-angle changes under relative motion. Let $D(\cdot)$ denote the target detector and $\Pi(\cdot)$ select the detected bbox of $\mathcal{V}$. At frame k, the attacked and clean boxes are $\hat{\mathbf{b}}_k = \Pi(D(\tilde{\mathcal{X}}_k))$ and $\mathbf{b}_k = \Pi(D(\mathcal{X}_k))$, with centers $\hat{\mathbf{c}}_k$ and $\mathbf{c}_k$. The attack aims to induce temporally coherent, progressively larger displacements of $\hat{\mathbf{c}}_k$ across $\mathcal{S}$ while keeping $\hat{\mathbf{b}}_k$ detectable.

4.2 Attack Design Overview

Inspired by viewing-angle variation naturally induced by relative motion in driving scenarios, we design a static physical camouflage attack that exploits this variation to produce temporally coherent 3D detection bias. As illustrated in Fig. 2, the design follows three stages: *Attack Scenario Data Preparation*, *Differentiable Camouflage Rendering*, and *Planning-Guided Attack Optimization*.

4.3 Attack Scenario Data Preparation and Camouflage Rendering

Optimizing an adversarial camouflage needs training data. To ensure the effectiveness of the attack, we employ a two-stage filtering process to select the valid consecutive sequence data $\mathcal{S}$ that has attack potentials, from an attacker-collected scenario dataset $\mathcal{D}$.

We first apply an **Attack Feasibility Filter (AFF)** to select an original sequence $\mathcal{O}$ where the target vehicle $\mathcal{V}$ remains visible, ahead of the ego vehicle, and relevant to its future path without already intruding into it, so that the attack has the potential to trigger hazardous planning responses by introducing detection displacement into a previously safe but highly relevant region of the ego-vehicle trajectory. **Viewing-Angle Filter (VAF)** then extracts a valid consecutive K-frame sequence $\mathcal{S}$ with sufficient viewing-angle variation between $\mathcal{X}_1$ and $\mathcal{X}_K$. Unlike prior attacks that suppress such variation, we favor larger variation so that the same δ can induce view-dependent responses and coherent motion illusions. After obtaining $\mathcal{S}$, we finally render δ onto the target vehicle under the corresponding camera geometry and visibility constraints to produce $\tilde{\mathcal{X}}_k$ for optimization.

4.4 Planning-Guided Attack Optimization

This stage optimizes δ by searching for a planning-guided displacement target and realizing it through detector-level optimization. A surrogate downstream pipeline determines the target direction and step size, while optimization requires only the target detector.

Optimal Target Search. We search for the target displacement pattern over candidate direction set $\mathcal{U}$ and step-size set $\mathcal{D}_s$. Let $\Gamma(\mathbf{u}, s)$ be the fake prediction under direction $\mathbf{u}$ and step size s, and $\mathcal{E}_{\text{plan}}(\cdot)$ its induced planning error. The optimal target is defined as

$$(\mathbf{u}^\star, s^\star) = \arg \max_{\mathbf{u}\in\mathcal{U},\, s\in\mathcal{D}_s} \mathcal{E}_{\text{plan}}\left(\Gamma(\mathbf{u}, s)\right). \tag{1}$$

Since δ is shared within a training group, we average the searched targets and obtain the group-level attack target $(\bar{\mathbf{u}}, \bar{s})$. The target-direction bbox displacement at frame k is

$$d_k = \left(\hat{\mathbf{c}}_k - \mathbf{c}_k\right)^\top \bar{\mathbf{u}}, \qquad k = 1, \ldots, K. \tag{2}$$

The planning-guided optimization target is to maximize the overall target-direction displacement while enforcing step consistency across consecutive frames:

$$\max_{\delta} \sum_{k=1}^{K} d_k - \lambda \sum_{k=1}^{K-1} \left(d_{k+1} - d_k - \bar{s}\right)^2. \tag{3}$$

Optimization Objectives. We implement the planning-guided target using move and progress losses, and regularize the optimization with fidelity and style losses. For physical robustness, we apply 3D Expectation over Transformation (EoT) (Athalye et al., 2018) with random yaw, translation, depth-ratio, and scale perturbations. The losses are defined below.

1. *Move loss.* The move loss drives the target 3D bbox center along the planning-guided direction $\bar{\mathbf{u}}$, i.e., maximizing the target-direction displacement d_k defined above.
2. *Progress loss.* The progress loss encourages progressive, spatio-temporally consistent bbox movement by matching consecutive displacement increments to $\bar{s}$.
3. *Fidelity loss.* The fidelity loss preserves detection credibility by penalizing deviations in confidence, size, and yaw, preventing target disappearance or severe box distortion. Let $\mathbf{r}$ and $\hat{\mathbf{r}}$ be the clean and attacked fidelity attribute vectors.
4. *Style loss.* The style loss $\mathcal{L}_{\text{style}}$ combines total variation loss $\mathcal{L}_{\text{tv}}$ to suppress high-frequency noise and non-printability-score loss $\mathcal{L}_{\text{nps}}$ to constrain colors to printable ranges.

$$\mathcal{L}_{\text{move}} = -\sum_{k=1}^{K} d_k, \qquad \mathcal{L}_{\text{prog}} = \sum_{k=1}^{K-1} \left(d_{k+1} - d_k - \bar{s}\right)^2, \qquad \mathcal{L}_{\text{fid}} = \left\|\hat{\mathbf{r}} - \mathbf{r}\right\|. \tag{4}$$

5 Evaluation

5.1 Evaluation Setups

Datasets. We evaluate our attack on nuScenes dataset (Caesar et al., 2020) using 220 attack scenarios, each with a contiguous sequence of camera feeds and annotations, across combinations of target vehicle type, relative position, and driving direction. Following prior attacks (Jia et al., 2020; Muller et al., 2022; Ma et al., 2025; Lou et al., 2024), we set $K = 3$ in the

main experiments, and also evaluate $K = 2, 4$ in Appendix D. Specifically, the target vehicle may be an SUV, sedan, or van, located in either the front-left (*L*) or front-right (*R*) region of the ego vehicle, and traveling in either the same (*S*) or opposite (*O*) direction. We denote each category as *TYPE-POS-DIR*; for example, *SUV-L-S* represents an SUV located in the front-left region of the ego vehicle and driving in the same direction. Such categories naturally involve attack scenarios in Fig. 1.

Evaluation settings. We evaluate two settings: *cross-validation* and *scenario-specific* attack. For the two primary categories, *SUV-R-S* and *SEDAN-R-S*, we perform 5-fold cross-validation, optimizing one shared camouflage on 4/5 of the scenarios and evaluating it on the remaining 1/5. This evaluates generalization to unseen scenarios within the same category. For the remaining categories with limited scenarios, we use the scenario-specific setting. For each case, one shared camouflage is optimized and evaluated on all available scenarios.

AD Models. We employ BEVDet (Huang et al., 2021), BEVDepth (Li et al., 2023), and Fast-BEV++ (Chen et al., 2025) as vision-based 3D detectors. To evaluate propagation, we use two downstream pipelines: Pipeline A comprises AB3DMOT (Weng et al., 2020), Trajectron++ (Salzmann et al., 2020), and an MPC-based planner (Chen et al., 2022); Pipeline B integrates CenterTrack (Zhou et al., 2020), HiVT (Zhou et al., 2022), and an FOT-based planner (Werling et al., 2010). We also evaluate our attack on the end-to-end model ST-P3 (Hu et al., 2022).

Metrics. We evaluate the attack at the perception, prediction, and planning levels using the metrics:

- *Frame-wise bbox displacement (d_1, d_2, d_3)* denotes the 3D bounding-box center displacement at the three attack frames, measured in meters along the split-specific average optimal direction $\bar{\mathbf{u}}$ in Eq. 2, which is estimated from each training split and fixed for its corresponding test split.
- *Progressive Displacement Rate (PDR)* measures the percentage of samples satisfying $d_1 < d_2 < d_3$, indicating temporally consistent bbox displacement across attack frames.
- *Average Prediction Error (APE)* measures the average deviation between the clean and attacked predicted trajectories of the target vehicle.
- *Minimum Trajectory Distance (MTD)* measures the minimum distance between the attacked predicted trajectory of the target vehicle and the victim AV's original planned trajectory. Smaller MTD indicates stronger planning interference.
- *Maximum Braking Deceleration (MBD)* measures the maximum braking deceleration triggered by the victim AV during the three attack frames.
- *Attack Success Rate (ASR)* is the percentage of successful attacks. An attack is considered successful if the victim AV triggers hard braking within the three attack frames, defined as $\mathrm{MBD} \geq 3.0\,\mathrm{m/s^2}$.

Implementation Details. We optimize the camouflage with Adam (Kingma & Ba, 2015) at a learning rate of 0.01 and 1024 $\times$ 1024 texture resolution, with PyTorch3D (Ravi et al., 2020) for differentiable rendering and SAM2 (Ravi et al., 2025) for vehicle segmentation. We run 3000 optimization steps for cross-validation and 500 for scenario-specific attacks. The attachable region covers the vehicle body excluding tires, windows, and mirrors. $\mathcal{U}$ contains 100 directions over 180°, and $\mathcal{D}_s$ 19 step sizes from 0.1 to 1.0 m at 0.05 m intervals. All experiments are conducted on a single NVIDIA H100.

5.2 Attack Effectiveness

Tab. 2 reports the main attack results on the two primary cases under 5-fold cross-validation. The proposed attack achieves up to 0.68 m progressive 3D bbox displacement across three detectors and triggers hard braking with up to 64.4% ASR through two downstream pipelines. All results in Tab. 2 are averaged over three random seeds; detailed repeated-run statistics are provided in Appendix E.

The attack induces temporally coherent perception bias. The results confirm that a static camouflage can reliably manipulate the target 3D bbox without dynamic patterns or online intervention. The average PDR reaches 73.7%, indicating that the induced bbox displacement is often temporally progressive rather than a single-frame perturbation. This supports our goal of generating coherent motion bias from frame-to-frame viewing-angle variation.

The perception bias propagates to downstream prediction and planning. Under Pipeline A/B, the attack produces average APE of 2.86/3.16 m, showing substantial deviation from the clean target-vehicle prediction. The average MTD of 1.68/1.48 m further indicate that the attacked prediction approaches the victim AV's original planned path, creating a safety-critical planning conflict. As

Table 2: Main attack results on two main scenarios and averaged over three random seeds. Values separated by '/' correspond to Pipeline A/B.

Scenario	Detector	3D BBox Displacement				Downstream Propagation			
		d_1 (m)	d_2 (m)	d_3 (m)	PDR (%)	APE (m)	MTD (m)	MBD (m/s^2)	ASR (%)
SUV-R-S	BEVDet	**0.15**	**0.35**	**0.68**	73.9	**2.97**/2.90	**1.46/1.29**	**3.52/4.14**	50.5/60.5
	BEVDepth	0.13	0.30	0.48	**74.5**	2.32/2.70	1.95/1.86	3.23/4.07	52.8/**61.1**
	FastBEV++	0.11	0.32	0.47	73.9	2.88/**3.23**	2.15/1.97	3.14/3.61	**55.0**/57.8
SEDAN-R-S	BEVDet	**0.14**	**0.32**	**0.65**	75.0	**3.65/3.81**	**1.04/0.90**	**3.90/4.84**	**58.3/64.4**
	BEVDepth	0.08	0.20	0.42	67.8	2.15/2.89	2.09/1.52	3.38/3.91	54.4/56.1
	FastBEV++	0.10	0.31	0.50	**77.2**	3.22/3.40	1.42/1.34	3.32/4.42	57.8/61.1
Average		0.12	0.30	0.53	73.7	2.86/3.16	1.68/1.48	3.42/4.16	54.8/60.2

Table 3: Specific-attack results on the cases with limited valid scenarios under Pipeline A.

Scenario	d_3 (m)	PDR (%)	ASR (%)	Succ./Total
SEDAN-L-O	0.85	73.7	68.4	13/19
SEDAN-L-S	0.79	84.6	76.9	10/13
SEDAN-R-O	0.83	85.7	71.4	10/14
SUV-L-O	0.84	87.5	75.0	6/8
SUV-L-S	**0.96**	77.8	66.7	6/9
SUV-R-O	0.91	83.3	66.7	4/6
VAN-L-O	0.91	**100.0**	75.0	3/4
VAN-L-S	0.94	92.3	84.6	11/13
VAN-R-O	0.93	83.3	83.3	5/6
VAN-R-S	0.93	87.5	**87.5**	**7/8**

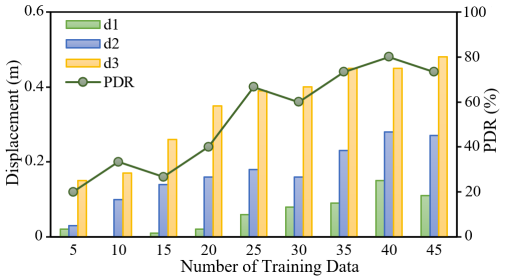


Figure 3: Effect of the number of training scenarios on cross-scenario generalization.

a result, the attack achieves average hard-braking ASR of 54.8%/60.2% under Pipeline A/B. On ST-P3 (Hu et al., 2022), the attack yields a 0.58 m average final displacement and a 32.5% ASR, with further results reported in Appendix H. We also achieve a **5.3× larger** d_3 and **4.3× higher** ASR than the strongest baseline; details are provided in Appendix G.

Successful attacks require sufficient cross-scenario training diversity. For the two largest cases, we fix 15 samples as the test set and vary the number of training samples from 5 to 45. As shown in Fig. 3, using too few training samples leads to unstable test performance, indicating underfitting of the camouflage. When the number of training samples reaches around 35-40, the d_3 and PDR become more stable, suggesting that cross-scenario generalization requires sufficient training samples.

The attack remains feasible across diverse traffic configurations. For the remaining cases, limited and uneven sample counts preclude train/test generalization evaluation. We therefore use a scenario-specific setting, optimizing one shared camouflage over all available samples in each case. As shown in Tab. 3, the attack remains effective across different vehicle types, relative positions, and directions, with an average d_3 of 0.89 m, PDR of 85.6%, and ASR of 75.6%.

5.3 Impact of Scenario Factors

Viewing-angle variation dominates under cross-validation. Fig. 4 analyzes five scenario factors: relative speed, effective projection area, distance, illumination and viewing angle. Since real-world data lacks strictly controlled variables and factors may be correlated, we use boxplots to compare the three-frame average displacement $\bar{d} = (d_1 + d_2 + d_3)/3$ across factor groups. Under cross-validation, scene variation and shared-camouflage generalization partially weaken factor effects. Nevertheless, larger viewing-angle variation yields stronger displacement, consistent with our core attack mechanism. Robustness to adverse weather and imaging perturbations is evaluated in Appendix C.

Scenario-specific optimization strengthens factor effects. Under the scenario-specific attack setting, one shared camouflage per case makes factor effects more pronounced. The attack is stronger with closer distance, brighter illumination, lower relative speed, larger projection area, and greater viewing-angle variation. Consistent with our design, larger projection and closer distance increase camouflage influence, brighter illumination improves visibility, lower relative speed smooths frame-to-frame changes, and larger viewing-angle variation amplifies view-dependent appearance evolution.

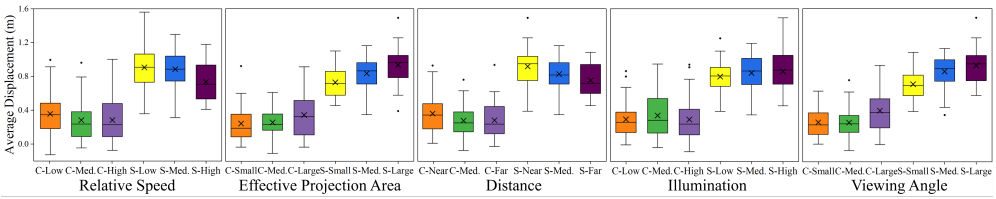


Figure 4: Impact of scenario factors—distribution of the three-frame average displacement $\bar{d}$ across different factor groups. C/S: evaluation settings cross-validation/scenario-specific.

Table 4: Cross-model transferability; reported 3D bbox displacement $d_1/d_2/d_3$ (m) in three frames.

Source Detector	Target Detector		
	BEVDet	BEVDepth	FastBEV++
BEVDet	**0.15/0.34/0.67**	-0.09/-0.08/0.01	-0.03/0.06/0.05
BEVDepth	-0.06/-0.02/-0.09	**0.11/0.25/0.45**	-0.14/0.02/0.13
FastBEV++	0.04/0.05/0.11	0.04/0.04/0.10	**0.11/0.32/0.49**

Table 5: Cross-vehicle-type transferability; reported 3D bbox displacement $d_1/d_2/d_3$ (m) in three frames.

Source Vehicle Type	Target Vehicle Type	
	SUV	Sedan
SUV	**0.13/0.32/0.54**	0.07/0.22/0.34
Sedan	0.09/0.23/0.35	**0.11/0.28/0.52**

5.4 Transferability

Cross-model transferability is challenging. Tab. 4 reports the cross-model transferability results. Camouflage optimized on one detector generally fails to preserve strong displacement on another detector, with the best transferred final-frame displacement reaching only 0.13 m. This limited transferability likely stems from detector-specific image-to-BEV transformations, feature aggregation strategies, and 3D box regression mechanisms, highlighting the fundamental challenge of inducing subtle box displacements under black-box settings. Ensemble optimization against two source detectors improves the displacement on the held-out detector by 0.09 m on average and up to 0.19 m, but overall transferability remains limited; further analysis is provided in Appendix F.

Cross-vehicle-type transfer remains effective. We further evaluate transferability across target-vehicle types. For each detector, we optimize the camouflage on one primary vehicle-type case and directly evaluate it on another after adapting the texture to the target vehicle body size and attachable surface. Tab. 5 shows that the transferred camouflage still induces non-trivial 3D bbox displacement in consecutive frames when attached to another vehicle in another scenario.

5.5 Ablation Study

Tab. 6 reports the ablation results on the two primary cases using BEVDet as the victim detector. We use Coefficient of Variation (CV) to measure displacement stability across perturbation conditions and Box Fidelity Score (BFS) to measure similarity between attacked and clean boxes in target-vehicle confidence and geometric attributes, where BFS ranges from 0 to 1 and 1 indicates identical fidelity. (1) **Movement and progression objectives are essential.** Removing $\mathcal{L}_{\text{move}}$ sharply reduces displacement, while removing $\mathcal{L}_{\text{prog}}$ mainly degrades PDR, showing that the two losses respectively drive bbox shifts and progressive temporal movement. (2) **VAF and EoT improve robustness and stability.** Removing VAF degrades overall performance, confirming the importance of explicitly exploiting viewing-angle variation, while removing EoT increases CV and weakens the overall results. (3) **Fidelity constraints stabilize the attack.** Removing $\mathcal{L}_{\text{fid}}$ sharply reduces BFS and also lowers ASR, indicating that preserving detection confidence and box geometry is important for maintaining downstream results. The effect of $\mathcal{L}_{\text{style}}$ is shown in Fig. 5. We further evaluate camouflage-area and localized-patch sensitivity in Appendix B, where reduced area consistently weakens effectiveness.

5.6 Dataset, Simulation, and Physical Case Studies

We present a representative case study of the full-pipeline attack in nuScenes in Fig. 6. The static camouflage exhibits view-dependent appearance changes under relative ego-target motion, causing progressive 3D bbox displacement of $d_1/d_2/d_3$ = 0.31/0.63/0.91 m, making the target appear to move toward the ego driving region, with an APE of 2.82 m and an MTD of 0.18 m, ultimately triggering hard braking by the ego vehicle. This confirms that our attack is not a single-frame perception



(a) w/ $\mathcal{L}_{\text{style}}$ (b) w/o $\mathcal{L}_{\text{style}}$

Figure 5: Ablation study of $\mathcal{L}_{\text{style}}$.

Table 6: Ablation study of key attack components. ASR is reported in the format Pipeline A/B.

Module	d_3 (m) ↑	CV ↓	PDR (%) ↑	BFS ↑	ASR (%) ↑
Full	**0.67**	0.52	**74.5**	0.83	**54.5/62.5**
w/o VAF	0.48	0.56	50.8	0.82	35.0/51.7
w/o EoT	0.43	0.77	54.2	0.76	46.7/48.3
w/o $\mathcal{L}_{\text{move}}$	0.12	1.05	51.7	**0.93**	15.8/23.3
w/o $\mathcal{L}_{\text{prog}}$	0.51	**0.48**	58.3	0.83	26.7/41.7
w/o $\mathcal{L}_{\text{fid}}$	0.64	2.04	65.0	0.60	49.2/57.5

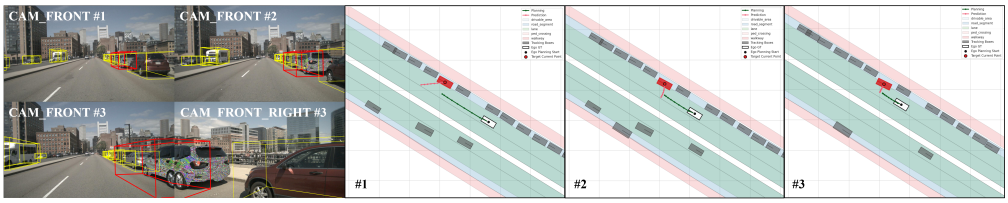


Figure 6: Representative case study of the attack results. Left: camera-view observations, including three consecutive frames; Right: the resulting BEV prediction and planning results.

perturbation; instead, it exploits viewing-angle variation to turn static camouflage into temporally coherent 3D bbox shifts that propagate through the downstream pipeline, causing unnecessary braking and supporting our motivation. The abandoned overtaking scenario is provided in Appendix A.

We further evaluate the attack through **closed-loop CARLA simulation** and a **physical print-and-recapture experiment**. In CARLA Town10HD, we deploy BEVDet with CenterTrack, HiVT, and PID control; the attack induces false lateral displacements of up to 0.96 m and triggers braking in three of four scenarios. We also print adversarial patches on a 1:15-scale vehicle and capture it with an approaching camera. The patches produce scaled detection-box deviations of up to 0.80 m in the optimization scene and 0.36 m in an unseen scene, demonstrating that the attack remains effective after physical printing and recapture. Full details are provided in Appendix I and Appendix J.

6 Discussion

Threats to validity. Our evaluation focuses on a limited set of perception models and autonomous driving systems. Although the attack transfers across different settings, the effectiveness may vary on different systems. In addition, our experiments are not conducted on commercial AD vehicles.

Mitigation. Potential defenses include multi-sensor fusion (Liu et al., 2023; Bai et al., 2022; Chen et al., 2023), temporal and physical consistency reasoning (Man et al., 2023; Yu et al., 2024), and adversarial patch or camouflage defenses (Liu et al., 2022; Kim et al., 2022; Jing et al., 2024). However, prior work shows that fusion systems remain vulnerable to sensor attacks (Cheng et al., 2024; Cao et al., 2021), and these approaches mitigate but do not fundamentally eliminate the vulnerability exploited in this work.

7 Conclusion

We present a view-induced trajectory manipulation attack that exploits natural viewing-angle variation to transform a static adversarial camouflage into temporally coherent motion deception. By inducing progressive 3D bounding-box displacement across consecutive frames, the attack propagates through downstream tracking, prediction, and planning modules, leading to harmful driving behaviors such as abandoned overtaking and unnecessary hard braking. Our results highlight viewing-angle variation as a new attack surface for autonomous driving systems and demonstrate the importance of evaluating security beyond single-frame perception robustness.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, language polishing, and LaTeX formatting. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work investigates potential security vulnerabilities in vision-based autonomous driving systems. All experiments are conducted using public driving data, open-source algorithms, controlled simulation, and a scaled physical vehicle model. We do not attack deployed autonomous vehicles or operational driving systems, and no real-world harm is caused during the study. Although the proposed attack may have dual-use implications, the goal of this research is to reveal an overlooked attack surface and motivate the development of more secure and robust autonomous-driving systems. We therefore intend these findings to support security research and defensive system design.

REPRODUCIBILITY STATEMENT

We provide the experimental settings, datasets, model configurations, evaluation metrics, and implementation details in Sec. 5.1. Additional experimental protocols, robustness evaluations, ablation studies, repeated-run statistics, and downstream evaluations are provided in the appendices. These details are intended to support reproduction of the reported results.

REFERENCES

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018.

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

Baidu. Baidu Apollo. `https://apollo.baidu.com/`, 2022.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11 (2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL `https://www.mdpi.com/2078-2489/11/2/125`.

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*, 2020.

Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE symposium on security and privacy (SP)*, pp. 176–194. IEEE, 2021.

Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition Workshops*, pp. 172–181, 2023.

Yuanpeng Chen, Hui Song, Wei Tao, ShanHui Mo, Shuang Zhang, Xiao Hua, and TianKun Zhao. Fastbev++: Fast by algorithm, deployable by design. *arXiv preprint arXiv:2512.08237*, 2025.

Yuxiao Chen, Ugo Rosolia, Wyatt Ubellacker, Noel Csomay-Shanklin, and Aaron Ames. Interactive multi-modal motion planning with branch model predictive control. *IEEE Robotics and Automation Letters*, 2022.

Zhiyuan Cheng, Hongjun Choi, Shiwei Feng, James Liang, Guanhong Tao, Dongfang Liu, Michael Zuzak, and Xiangyu Zhang. Fusion is not enough: Single modal attacks on fusion models for 3d object detection. In *International Conference on Learning Representations*, 2024.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 2022.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L. Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Xiaoyu Ji, Yushi Cheng, Yuepeng Zhang, Kai Wang, Chen Yan, Wenyuan Xu, and Kevin Fu. Poltergeist: Acoustic adversarial machine learning against cameras and computer vision. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 160–175, 2021.

Yunhan Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations*, 2020.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

Lihua Jing, Rui Wang, Wenqi Ren, Xin Dong, and Cong Zou. Pad: Patch-agnostic defense against adversarial patch attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24472–24481, 2024.

Taeheon Kim, Youngjoon Yu, and Yong Man Ro. Defending physical adversarial attack on object detection via adversarial patch-feature energy. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1905–1913, 2022.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 1477–1485, 2023.

Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14973–14982, 2022.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. IEEE, 2023.

Tianrui Lou, Xiaojun Jia, Siyuan Liang, Jiawei Liang, Ming Zhang, Yanjun Xiao, and Xiaochun Cao. 3d gaussian splatting driven multi-view robust physical adversarial camouflage generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28752–28762, 2025.

Yang Lou, Yi Zhu, Qun Song, Rui Tan, Chunming Qiao, Wei-Bin Lee, and Jianping Wang. A first physical-world trajectory prediction attack via lidar-induced deceptions in autonomous driving. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 6291–6308, 2024.

Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. Wip: Towards the practicality of the adversarial attack on object tracking in autonomous driving. In *ISOC Symposium on Vehicle Security and Privacy*, 2023.

Chen Ma, Ningfei Wang, Zhengyu Zhao, Qian Wang, Qi Alfred Chen, and Chao Shen. Controlloc: Physical-world hijacking attack on camera-based perception in autonomous driving. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 738–752, 2025.

Yanmao Man, Ming Li, and Ryan Gerdes. Ghostimage: Remote perception domain attacks against camera-based image classification systems. In *Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses (USENIX RAID 2020)*, 2020.

Yanmao Man, Raymond Muller, Ming Li, Z Berkay Celik, and Ryan Gerdes. That person moves like a car: Misclassification attack detection for autonomous systems using spatiotemporal consistency. In *32nd USENIX Security Symposium*, pp. 6929–6946, 2023.

Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8):1909–1963, 2023.

Raymond Muller, Yanmao Man, Z Berkay Celik, Ming Li, and Ryan Gerdes. Physical hijacking attacks against object trackers. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2309–2322, 2022.

Raymond Muller, Ruoyu Song, Chenyi Wang, Yuxia Zhan, Jean-Philippe Monteuuis, Yanmao Man, Ming Li, Ryan Gerdes, Jonathan Petit, and Z. Berkay Celik. Investigating physical latency attacks against camera-based perception. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 4588–4605, 2025.

Ben Nassi, Yisroel Mirsky, Dudi Nassi, Raz Ben-Netanel, Oleg Drokin, and Yuval Elovici. Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 293–308, 2020.

NVIDIA. Nvidia-alpamayo. `https://www.nvidia.cn/solutions/autonomous-vehicles/alpamayo/`, 2026.

Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10142–10162, 2021.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision (ECCV)*, 2020.

Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Dta: Physical camouflage attacks using differentiable transformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15305–15314, 2022.

Naufal Suryanto, Yongsu Kim, Harashta Tatimma Larasati, Hyoeun Kang, Thi-Thu-Huong Le, Yoonyoung Hong, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Active: Towards highly transferable 3d physical camouflage for universal and robust vehicle evasion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4305–4314, 2023.

Tesla. Tesla Full Self-Driving. `https://www.tesla.com/fsd`, 2026.

Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Zhiqiang Gong, Xiaoya Zhang, Wen Yao, and Xiaoqian Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 2414–2422, 2022.

Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8565–8574, 2021.

Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4412–4423, 2023.

Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

Moritz Werling, Julius Ziegler, Sören Kammel, and Sebastian Thrun. Optimal trajectory generation for dynamic street scenarios in a frenet frame. In *2010 IEEE international conference on robotics and automation*, pp. 987–993. IEEE, 2010.

Yi Yu, Weizhen Han, Libing Wu, Bingyi Liu, Enshu Wang, and Zhuangzhuang Zhang. Enduring, efficient and robust trajectory prediction attack in autonomous driving via optimization-driven multi-frame perturbation framework. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17229–17238, 2025.

Zhiyuan Yu, Ao Li, Ruoyao Wen, Yijia Chen, and Ning Zhang. Physense: Defending physically realizable attacks for autonomous systems via consistency reasoning. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3853–3867, 2024.

Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15159–15168, 2022.

Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *European Conference on Computer Vision*, 2020.

Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Wenjun Zhu, Xiaoyu Ji, Yushi Cheng, Shibo Zhang, and Wenyuan Xu. Tpatch: A triggered physical adversarial patch. In *32nd USENIX Security Symposium*, pp. 661–678, 2023.

A Additional Case Study

In addition to the hard-braking case in the main text, we provide an additional case study of abandoned overtaking in Fig. 7. In the clean scene, the ego vehicle is planning to overtake the target vehicle and continue along the original driving region. Under attack, the induced 3D bbox displacement increases across frames with $d_1/d_2/d_3 = 0.42/0.77/1.16$ m, shifting the predicted target trajectory closer to the ego driving region and reducing the MTD to 0.05 m, with an APE of 2.80 m. Consequently, the victim AV abandons the overtaking maneuver.

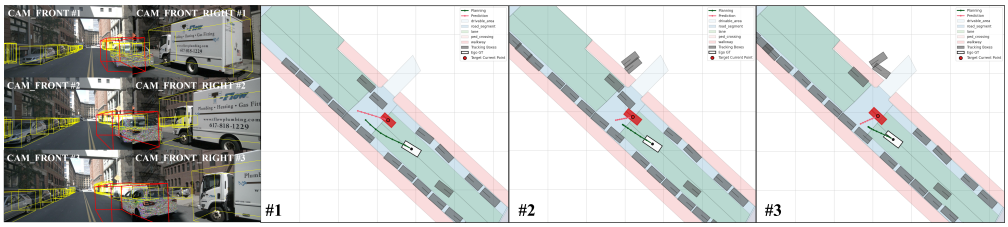


Figure 7: Additional case study of abandoned overtaking. In the clean scene, the target vehicle is preparing to turn right, and the ego vehicle plans to overtake it. Under attack, the induced 3D bbox displacement makes the victim AV perceive the target vehicle as drifting left toward the ego driving region, reducing the clearance required for overtaking and causing the planner to abandon the original overtaking maneuver. Left: camera-view observations, including three consecutive frames; Right: the resulting BEV prediction and planning results.

B Attack Performance under Area Constraints

Our default attachable region excludes the tires, windows, and mirrors and is restricted to paintable body panels, following prior camouflage attacks (Wang et al., 2022; Suryanto et al., 2022; 2023). To evaluate area sensitivity, we use the same cross-validation setting as the main experiments with 120 samples, including 60 *SUV-R-S* and 60 *SEDAN-R-S* samples across three detectors and two downstream pipelines. We reduce the original attachable region to 75%, 50%, and 25%, and additionally evaluate rectangular patches placed on the vehicle side and rear with sizes of 80%, 60%, and 40% of the corresponding vehicle dimensions. Examples of the evaluated camouflage-area constraints are shown in Fig. 8.

Table 7: Attack performance under different camouflage-area constraints. d_1 (m), d_2 (m), d_3 (m), PDR (%), APE (m), MTD (m), MBD (m/s^2), and ASR (%) are reported, where APE, MTD, MBD, and ASR correspond to Pipelines A/B.

Setting	$d_1 \uparrow$	$d_2 \uparrow$	$d_3 \uparrow$	PDR ↑	APE ↑	MTD ↓	MBD ↑	ASR ↑
Camouflage 100%	0.12	0.30	0.53	73.7	2.86/3.16	1.68/1.48	3.42/4.16	54.8/60.2
Camouflage 75%	0.08	0.15	0.41	56.7	2.04/2.47	1.87/1.75	2.67/3.71	43.1/51.4
Camouflage 50%	0.06	0.12	0.32	50.8	1.93/2.29	3.56/3.11	1.28/1.49	29.4/38.6
Camouflage 25%	0.02	0.07	0.13	31.7	1.17/1.55	4.25/3.96	0.91/1.03	11.7/19.7
Patch 80%	0.07	0.18	0.45	51.7	2.52/2.89	1.71/1.67	2.86/3.43	46.9/52.5
Patch 60%	0.05	0.11	0.42	47.2	2.08/2.35	2.53/2.41	2.11/2.73	36.1/40.3
Patch 40%	0.02	0.06	0.17	33.8	1.59/1.98	3.32/3.09	1.36/1.89	23.6/25.9

As shown in Tab. 7, as the camouflage coverage decreases from 100% to 25%, the final-frame displacement decreases from 0.53 m to 0.13 m, while ASR decreases from 54.8%/60.2% to 11.7%/19.7% under Pipelines A/B. Localized patches exhibit a similar trend, indicating that a larger available attack area generally enables stronger and more stable manipulation.

C Robustness to Physical Factors

As shown in Fig. 9, we evaluate physical robustness from two complementary aspects: naturally occurring adverse weather in nuScenes and controlled imaging perturbations. Under the same 120-



(a) Camouflage 100% (b) Camouflage 75% (c) Camouflage 50% (d) Camouflage 25%



(e) Patch 80% (f) Patch 60% (g) Patch 40%

Figure 8: Visualization of different camouflage-area constraints. The top row shows full-surface camouflage with 100%, 75%, 50%, and 25% of the original attachable area. The bottom row shows rectangular patches with 80%, 60%, and 40% of the corresponding vehicle dimensions.



(a) Normal weather (b) Increased illumination (c) Reduced illumination

(d) Motion blur (e) Fog (f) Rain spatter

Figure 9: Visualization of the evaluated physical factors. The first panel shows a naturally occurring normal-weather sample from nuScenes. The remaining panels show controlled imaging perturbations applied to attacked images, including increased illumination, reduced illumination, motion blur, fog, and rain spatter.

sample cross-validation setting as the main experiments, samples are selected solely based on attack feasibility rather than weather conditions, including 79 normal-weather and 41 rainy/foggy samples. Compared with normal weather, rainy/foggy conditions increase the final-frame displacement by 11.8%, reduce PDR by 11.6%, decrease the average MTD by 3.30%, and increase the average MBD by 3.92%. These results indicate that the attack remains effective on adverse-weather samples contained in the dataset.

We also apply consistent illumination, fog, motion-blur, and rain-spatter transformations to the complete images of the same 120 samples using Albumentations (Buslaev et al., 2020). Relative to the original setting, increased illumination, reduced illumination, motion blur, and fog reduce the final-frame displacement by 4.5%, 1.6%, 8.3%, and 20.9%, respectively, whereas rain spatter increases it by 14.0%, although the progression during the first two frames becomes less stable. Overall, the attack remains effective under common environmental and imaging perturbations, with fog causing the largest degradation among the evaluated conditions.

D EFFECT OF K

We evaluate $K = 2$, 3, and 4 under the same cross-validation setting as the main experiments, with results averaged across the two primary scenarios and three detectors. As shown in Tab. 8, $K = 2$ achieves the highest PDR of 89.2% because progressive displacement is easier to maintain over fewer frames, but its shorter attack duration results in lower downstream impact, with ASRs of 44.3%/46.9% under Pipelines A/B. Increasing to $K = 3$ improves the downstream effect, achieving ASRs of 54.8%/60.2%. For $K = 4$, the extended attack duration maintains comparable ASRs of 54.9%/59.6%, while PDR decreases to 54.2%, indicating that enforcing progressive displacement becomes more difficult over longer sequences. These results support $K = 3$ as a practical balance between temporal progression and downstream attack effectiveness.

Table 8: Attack performance with different numbers of attack frames. d_1 (m), d_2 (m), d_3 (m), d_4 (m), PDR (%), APE (m), MTD (m), MBD (m/s^2), and ASR (%) are reported, where APE, MTD, MBD, and ASR correspond to Pipelines A/B.

K	$d_1 \uparrow$	$d_2 \uparrow$	$d_3 \uparrow$	$d_4 \uparrow$	PDR $\uparrow$	APE $\uparrow$	MTD $\downarrow$	MBD $\uparrow$	ASR $\uparrow$
2	0.26	0.57	–	–	89.2	2.11/2.54	2.37/2.05	2.68/3.01	44.3/46.9
3	0.12	0.30	0.53	–	73.7	2.86/3.16	1.68/1.48	3.42/4.16	54.8/60.2
4	0.02	0.14	0.34	0.50	54.2	2.33/2.79	1.36/1.29	2.89/4.10	54.9/59.6

E REPEATED RUNS AND STATISTICAL UNCERTAINTY

We repeat the complete cross-validation experiments using three random seeds while keeping all other experimental settings unchanged. The evaluation covers the same 120 samples, consisting of 60 SUV-R-S and 60 SEDAN-R-S samples, across BEVDet (Huang et al., 2021), BEVDepth (Li et al., 2023), and FastBEV++ (Chen et al., 2025). Each scenario–detector combination therefore contains 60 samples in each run. Table 9 reports the mean and sample standard deviation over the three runs.

Across the six scenario–detector combinations, the average standard deviations are 0.027 m for the final-frame displacement d_3, 5.61% for PDR, and 4.80%/3.04% for ASR in Pipelines A/B, respectively. The repeated runs consistently preserve the progressive 3D displacement and downstream attack effects across both scenarios and all three detectors, indicating that the main conclusions are stable across random seeds.

Table 9: Repeated-run results over three random seeds. Values are reported as mean $\pm$ sample standard deviation. ASR-A/B denote the attack success rates under Pipeline A/B.

Scenario	Detector	d_3 (m)	PDR (%)	ASR-A (%)	ASR-B (%)
SUV-R-S	BEVDet	0.677 ± 0.025	73.87 ± 4.18	50.53 ± 2.54	60.53 ± 2.54
	BEVDepth	0.477 ± 0.015	74.47 ± 6.93	52.77 ± 7.51	61.13 ± 4.18
	FastBEV++	0.470 ± 0.026	73.87 ± 4.18	55.00 ± 1.70	57.80 ± 1.91
SEDAN-R-S	BEVDet	0.647 ± 0.045	75.00 ± 5.00	58.33 ± 6.02	64.43 ± 1.96
	BEVDepth	0.420 ± 0.030	67.77 ± 9.16	54.43 ± 6.77	56.13 ± 5.10
	FastBEV++	0.503 ± 0.021	77.23 ± 4.23	57.77 ± 4.23	61.10 ± 2.55
Average Std.		0.027	5.61	4.80	3.04

F FURTHER ANALYSIS OF CROSS-MODEL TRANSFERABILITY

Ensemble optimization improves, but does not solve, black-box transferability. To further mitigate the cross-model discrepancy observed in Tab. 4, we jointly optimize the camouflage against two of the three 3D detectors and evaluate it on the third held-out detector using the same 120-sample cross-validation setting. As summarized in the main paper, ensemble optimization improves the final-frame displacement by 0.09 m on average and up to 0.19 m compared with single-source transfer.

Nevertheless, the transferred attack effect remains substantially weaker than the corresponding white-box performance, suggesting that multi-model optimization only partially alleviates detector-specific overfitting.

The optimization directions are highly model-specific. We further analyze the limited transferability from three complementary perspectives. First, we compute the cosine similarity between single-step attack gradients produced by the three detectors under identical input samples, texture initialization, and attack objectives. The pairwise similarities are only 0.029–0.054, showing that the local update directions are nearly orthogonal across models. Consequently, a texture update that is effective for one detector provides little aligned descent direction for another detector.

Second, we directly evaluate cross-model optimization by optimizing the camouflage on one detector and testing it on another. Among the six directed source-to-target detector pairs, four exhibit negative transfer gains. For the remaining two pairs, the transferred improvements retain only 3.8% and 13.2% of the corresponding white-box gains. This result further suggests that the effective optimization landscape differs substantially across detector architectures.

Third, we analyze the overlap of effective perturbation subspaces obtained from 100 random texture initializations. The overlap between detector pairs is consistently low, indicating that the perturbation directions that effectively induce the desired 3D displacement occupy largely different subspaces across the three detectors. Together with the gradient-similarity and cross-optimization results, this analysis suggests that gradient-based optimization for the proposed directional displacement objective is strongly model-dependent.

Why transferability is particularly challenging for our objective. This limitation is closely related to the attack objective considered in this work. Prior physical camouflage attacks such as FCA (Wang et al., 2022), DTA (Suryanto et al., 2022), and ACTIVE (Suryanto et al., 2023) mainly optimize objectives related to 2D detection confidence, objectness, or generic detection degradation. In contrast, our attack directly targets camera-based 3D detectors and requires the predicted 3D bounding box to move along a prescribed planning-relevant direction, with the displacement increasing across consecutive frames while the object remains successfully detected. The attack therefore depends not only on whether the detector is perturbed, but also on how model-specific visual features are mapped to 3D geometric attributes and how these attributes evolve across viewpoints.

Such a directional and temporally progressive objective imposes a substantially stronger structural constraint than reducing detection confidence alone. Differences in feature encoding, depth estimation, BEV construction, and 3D box regression can therefore lead to markedly different gradient directions and effective perturbation subspaces across detectors. The ensemble results indicate that multi-model optimization can partially alleviate this discrepancy, but strong black-box transfer across heterogeneous 3D detectors remains an open challenge. Improving this transferability, for example through stronger model ensembles or model-agnostic optimization objectives, is an important direction for future work.

G Comparison with Prior Baselines

We compare our method with three representative related attacks: the tracker-hijacking attack of Jia et al. (2020), AttrackZone (Muller et al., 2022), and ControlLoc (Ma et al., 2025). These methods primarily manipulate 2D detection or tracking outputs and typically rely on dynamically displayed or projected perturbations. In contrast, our method uses a single static camouflage to induce progressively increasing 3D bounding-box displacement along a planning-relevant direction by explicitly exploiting natural viewpoint evolution across frames. The temporal effect therefore arises from the interaction between static camouflage, vehicle geometry, and relative motion, rather than from a dynamically changing perturbation or an isolated temporal regularizer.

We evaluate all four methods using BEVDet, BEVDepth, and FastBEV++ under the same 120-sample cross-validation protocol, consisting of 60 SUV and 60 sedan samples, and report the same downstream metrics.

As shown in Tab. 10, compared with the strongest prior baseline for each metric, our method increases the final-frame displacement by 5.3× and achieves approximately 4.4×/4.2× higher ASR under Pipelines A/B. It also yields larger prediction errors and braking responses while reducing the

Table 10: Comparison with prior baselines under the same downstream evaluation protocol. Values separated by '/' correspond to Pipeline A/B.

Method	d_3 (m) ↑	APE (m) ↑	MTD (m) ↓	MBD (m/s^2) ↑	ASR (%) ↑
Jia et al.	0.10	1.48/2.00	3.31/3.50	0.88/1.86	12.5/13.3
ControlLoc	0.09	1.47/1.93	3.15/3.24	0.65/1.29	10.8/14.2
AttrackZone	0.04	0.94/1.78	3.59/3.68	0.15/0.93	3.33/6.67
Ours	**0.53**	**2.86/3.16**	**1.68/1.48**	**3.42/4.16**	**54.8/60.2**

minimum trajectory distance. These results show that exploiting viewpoint evolution to generate progressive 3D displacement produces substantially stronger downstream effects than directly adapting existing 2D detection- or tracking-oriented physical attacks.

H Evaluation on an End-to-End Driving Model

To examine whether the proposed attack depends on the explicit perception–tracking interface used in the modular pipelines, we further evaluate it on ST-P3 (Hu et al., 2022), an end-to-end vision-based autonomous-driving model with spatial-temporal feature learning. We use the same 120-sample split as in the main experiments, consisting of 60 SUV-R-S and 60 SEDAN-R-S samples.



(a) Clean output of ST-P3.



(b) Attacked output of ST-P3.

Figure 10: Planning outputs of ST-P3 in the clean and attacked settings. The proposed attack triggers hard braking.

Across the 120 samples, the attack achieves an MBD of 2.31 m/s^2 and an ASR of 32.5%. We additionally measure the L2 displacement between the attacked and clean planning trajectories. The average displacement is 0.40 m over all planning points and increases to 0.58 m at the final planning point, with maximum displacements of 3.49 m and 4.98 m, respectively. Fig. 10 provides representative examples of the resulting planning deviations.

These results show that the effect of static camouflage can propagate directly to the planning output of an end-to-end model. In the modular pipelines, progressive 3D bounding-box displacement provides an explicit and interpretable manifestation of the attack. The ST-P3 results indicate that this explicit interface is not necessary for the attack to influence planning: image-level perturbations can also propagate through implicit BEV and temporal representations and alter the final trajectory.

More broadly, VLA systems introduce substantially different visual representations, reasoning processes, and action-generation mechanisms from both modular pipelines and current end-to-end



(a) Left, same direction (b) Left, oncoming (c) Right, same direction (d) Intersection

Figure 11: Attacks in closed-loop CARLA simulation under four driving scenarios.

Table 11: Closed-loop results in CARLA (Town10HD).

Scenario	$d_1/d_2/d_3$ (m)	Closest appr. (m)	Ego speed (25 km/h $\rightarrow$)	Effect
Left, same direction	$0.04/0.18/0.50$	0.28	11 km/h	Emergency brake
Left, oncoming	$0.61/0.24/0.96$	0.68	11 km/h	Emergency brake
Right, same direction	$0.07/0.23/0.39$	0.71	11 km/h	Emergency brake
Intersection (85° turn)	$-0.06/0.24/0.36$	1.93	25 km/h	No brake

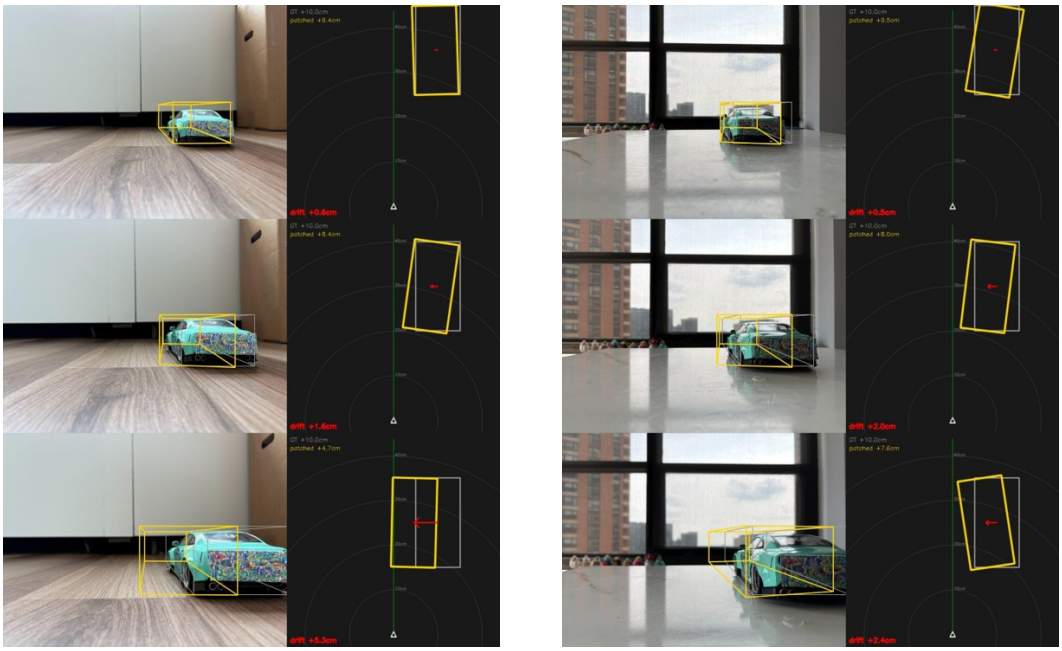
driving models. Evaluating effective and transferable physical attacks in this setting therefore remains an important direction for future work.

I Closed-Loop Simulation in CARLA

To evaluate the attack in a closed-loop driving setting, we port it to CARLA Town10HD. A static camouflage pattern is optimized over the body panels of a target vehicle and rendered in CARLA. Because CARLA does not provide a differentiable renderer, we construct a differentiable *paint-response* surrogate from renderings under different uniform paint colors. The surrogate maps the camouflage texture to its rendered appearance using fixed texture coordinates and piecewise-linear color-response curves, achieving a mean absolute error of approximately $5/255$ on held-out colors. A pretrained BEVDet detector is followed by ego-motion-compensated tracking, constant-velocity prediction with a 6-s horizon, and an ACC planner with PID-based emergency braking.

We evaluate four three-keyframe driving scenarios, as visualized in Fig.11, with frames spaced 0.5 s apart and the ego vehicle initially traveling at 25 km/h. The target vehicle is located to the front-left or front-right, traveling in the same or opposite direction, or turning at an intersection. For each scenario, the camouflage is optimized to induce a progressive lateral detection drift toward the ego lane across the three frames.

As summarized in Table 11, the attack induces sub-meter per-frame detection drifts that are aggregated by tracking and prediction into false cut-ins of up to 2.66 m. In the left-same-direction, left-oncoming, and right-same-direction scenarios, the predicted target trajectory enters the ego lane, with closest approaches of 0.28, 0.68, and 0.71 m, respectively. The planner consequently triggers emergency braking and reduces the ego speed from 25 km/h to approximately 11 km/h, whereas the corresponding clean scenes trigger no braking. In the intersection scenario, the closest approach remains outside the ego lane at 1.93 m, and no braking occurs. Overall, three of the four scenarios trigger braking,



(a) Known scene (b) Unknown scene

Figure 12: Attacks on a scaled car model using physically printed adversarial patches. Left: known scene, using the same scene as optimization. Right: unknown scene, using a different scene from optimization.

Table 12: Detection-box deviations in the physical print-and-recapture experiment over three key frames. Scaled deviations are reported according to the scale-model setting.

Scene	Real-world deviation (cm)	Scaled deviation (m)
Known	0.6 / 1.6 / 5.3	0.09 / 0.24 / 0.80
Unknown	0.5 / 2.0 / 2.4	0.08 / 0.30 / 0.36

demonstrating that small but temporally coherent perception errors can be amplified by tracking and prediction into safety-critical closed-loop behavior.

J Physical Print-and-Recapture Experiment

We further conduct a physical print-and-recapture experiment using a 1:15-scale vehicle model and a pretrained BEVDet detector. Images are captured using an iPhone 16, with the camera approaching the vehicle model from the rear to mimic the attack scenario considered in the paper. Two flat adversarial patches are placed on the rear and side body panels, with physical sizes of 9×3 cm and 6×4 cm, respectively.

The adversarial patches are optimized through a differentiable homography compositor and BEVDet to shift the detected 3D bounding box laterally toward the ego centerline. After optimization, the patches are printed at their exact physical sizes, attached to the scale model, and re-photographed. We measure the detection-box deviations across three key frames under two settings. In the known scene, optimization and physical evaluation use the same background. In the unknown scene, the printed attack is evaluated in a different setting from that used during optimization. Representative examples of the two physical evaluation settings are shown in Fig. 12.

As summarized in Tab. 12, the printed patches induce detection-box drifts across the three frames in both settings. In the known scene, the measured deviations are 0.6/1.6/5.3 cm on the scale model, corresponding to scaled deviations of 0.09/0.24/0.80 m. In the unknown scene, the deviations are

0.5/2.0/2.4 cm, corresponding to 0.08/0.30/0.36 m after scaling. These results show that detection-box drifts remain observable after physical printing and recapture, including when the attack is evaluated in a scene different from that used during optimization, validating the physical-world feasibility of the attack.

However, experiments using a pretrained detector, closed-loop simulation, and a scale vehicle model do not fully represent industrial autonomous-driving deployment. The real-world impact of the attack remains dependent on the specific system, sensor, and model configurations.