

Agent-OBJ: Prompt-Driven 3D Adversaries for Multi-modal Perception

Bing Li¹, Sean Du¹, Xuhong Ren^{2†}, Luqi Gong⁴, Wee Peng Tay¹, and Qing Guo^{2,3†}

¹ Nanyang Technological University, Singapore
`bing.li@ntu.edu.sg`

² NKIARI, Shenzhen Futian, China

³ AAIS & VCIP, Nankai University, China
`tsingguo@ieee.org`

⁴ Research Center for Space Computing System, Zhejiang Lab, China

Abstract. Modern autonomous driving systems increasingly rely on tightly coupled Camera–LiDAR fusion pipelines to achieve robust scene understanding. While this multi-modal redundancy is designed to counter independent single-sensor failures, it implicitly assumes that physical objects cannot simultaneously deceive both modalities. In this paper, we expose the vulnerability of this assumption by proposing Agent-OBJ, a generative framework that synthesizes physically plausible 3D adversaries via a prompt-driven agent. Unlike prior works that mainly focus on sensor-specific perturbations, our pipeline instantiates a physically plausible 3D adversary by generating a canonical pedestrian geometry from a base prompt and then modulating its shape via an additional semantic prompt, while controlling appearance with a pretrained multi-view personalization model conditioned on multi-view images and a style prompt to produce view-consistent appearance. To simultaneously compromise both 2D monocular and 3D fusion detectors, the agent dynamically optimizes the generated adversary using confidence-score feedback from both detectors. Meanwhile, to ensure stealthiness, we regularize the generated 3D point cloud by enforcing similarity to a pedestrian feature bank, aligning it with the distribution of natural objects and making it visually and geometrically indistinguishable from benign instances. Furthermore, we incorporate a multi-view consistency constraint during optimization to promote cross-view robustness, ensuring the adversary remains effective under diverse viewpoints in real-world driving scenes. Extensive experiments on nuScenes demonstrate high attack success rates against both monocular and fusion detectors while preserving strong perceptual plausibility.

Keywords: Adversarial Attacks, Camera–LiDAR Fusion, 3D Adversarial Objects

[†] Xuhong Ren and Qing Guo are Corresponding authors.

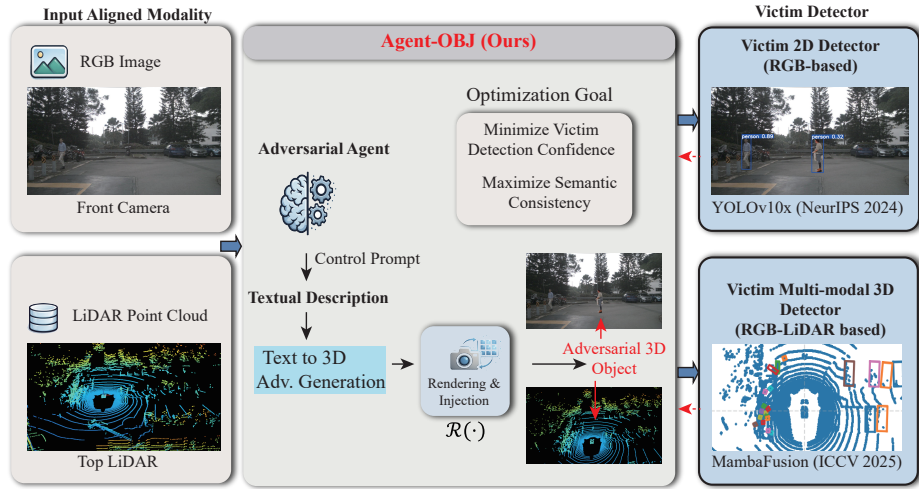


Fig. 1: Agent-OBj pipeline. An LLM agent edits semantic and style prompts to steer a text-to-3D generator, producing a physically plausible pedestrian-like object. The object is rendered into aligned camera–LiDAR views and evaluated by both a monocular 2D detector and a Camera–LiDAR fusion 3D detector using only black-box confidence feedback.

1 Introduction

Autonomous driving perception [18, 26, 31, 37, 48] increasingly relies on multi-sensor stacks, where monocular camera perception and Camera–LiDAR fusion are widely adopted to enhance robustness. By combining complementary sensing principles, fusion systems are expected to mitigate modality-specific noise and reduce the likelihood of catastrophic failures. A common implicit safety assumption behind such redundancy is that a single *physical* object is unlikely to simultaneously mislead multiple modalities in a consistent manner. In this work, we revisit this assumption and ask a fundamental question: can *object-level*, physically realizable adversaries remain effective across both monocular and fusion pipelines?

Realizing such object-level multi-modal adversaries is non-trivial due to three inherent constraints. First, the cross-modal signal gap: cameras and LiDAR capture fundamentally different signals (texture vs. geometry), making it difficult to find a single perturbation that simultaneously degrades both. Second, the dynamic viewing condition: autonomous driving involves drastic changes in distance and viewpoint; an adversary optimized for one pose often fails as the vehicle moves. Third, the black-box nature of production-grade stacks: real systems often involve non-differentiable preprocessing (e.g., NMS) and hide their internal gradients, necessitating a query-efficient optimization paradigm.

Most prior adversarial research targets a single sensing stream—either camera-only [2, 17, 20, 51, 52, 60] or LiDAR-only [5, 7, 22, 36, 38, 41, 62]—and often assumes white-box access (or differentiable surrogates) to optimize perturbations. Recent

3D object attacks [6, 24] further explore mesh- or shape-level manipulations, yet they typically rely on gradient-based optimization through renderers or model internals and are mainly evaluated against *single-modality* detectors; as a result, the optimized geometries may deviate from natural object priors or overfit to limited viewpoints. These limitations become particularly pronounced in realistic driving scenarios with large viewpoint/distance variations, and they do not naturally account for the cross-modal alignment and consistency mechanisms inherent to Camera–LiDAR fusion stacks. This gap motivates an object-level, black-box attack paradigm that can synthesize a *single* physically plausible 3D entity with controllable semantics and appearance, and optimize it to degrade both monocular and fusion detectors simultaneously. Crucially, such optimization must preserve physical plausibility and multi-view consistency, producing a stable adversary rather than a viewpoint-specific artifact.

Building on these insights, we propose **Agent-OBJ**, a prompt-driven generative framework that synthesizes physically plausible 3D adversaries and induces consistent cross-system failures in autonomous driving perception (Fig. 1). As illustrated, Agent-OBJ instantiates a canonical pedestrian geometry from a base prompt, modulates its shape via an additional semantic prompt, and controls appearance using a pretrained multi-view personalization model conditioned on multi-view images and a style prompt to produce view-consistent textures. It then performs score-based, gradient-free optimization in a strict black-box setting, using only joint confidence feedback from a monocular 2D detector and a Camera–LiDAR fusion 3D detector. To ensure stealthiness, we regularize the generated point cloud by encouraging similarity to a pedestrian feature bank, aligning it with the distribution of natural objects. Moreover, we incorporate a multi-view consistency constraint to promote cross-view robustness under diverse viewpoints. Extensive evaluations on nuScenes demonstrate that Agent-OBJ achieves strong and stable adversarial effects against both monocular and fusion detectors while remaining highly inconspicuous.

Our main contributions are four-fold:

- We formulate a practical *black-box, object-level* threat model for *unified* attacks on monocular 2D and Camera–LiDAR fusion 3D perception, emphasizing multi-view robustness under realistic driving viewpoints.
- We propose Agent-OBJ, a prompt-driven generative pipeline that decouples *semantic* (geometry) and *style* (appearance) control, enabling discrete, controllable exploration of physically plausible 3D adversarial objects.
- We introduce a joint optimization objective with naturalness and multi-view stability regularization, producing adversaries that are both effective across modalities and visually/geometrically plausible.
- We demonstrate strong attack success, transferability across detectors, and improved query efficiency on nuScenes under a fixed black-box query budget.

2 Related Work

Camera-based Attacks. Attacks on monocular perception typically rely on *pixel-level* perturbations [53, 55] or localized *adversarial patches* [3, 11]. Early physical-world studies highlight the importance of *transformation robustness* by optimizing over viewpoint, scale, illumination, and printing effects (e.g., EOT/RP2) [1, 12, 39]. Recent works further improve patch deployability for detection via triggerable/camouflaged designs and robustness-oriented training or decomposition strategies [2, 9, 17, 20, 51, 52, 60, 63]. Beyond detection, adversarial data augmentation has also been explored to improve robustness in low-data regimes such as few-shot classification [56].

Despite strong progress, these methods remain fundamentally constrained to *image-space* manipulations: they primarily compromise the camera branch and often exhibit limited effectiveness as *object-level* adversaries that generalize to multi-modal perception stacks. Beyond detection, related lines also study attacks on tracking and segmentation, but they similarly operate on camera-specific representations and typically do not address cross-modal consistency required by modern fusion systems.

LiDAR-based 3D Attacks. Threats to 3D perception span *digital* point-cloud manipulations [28, 49, 50] and *physical* interventions. The latter includes active sensor spoofing (e.g., laser injection and aiming) [5, 7, 22, 36, 38, 41, 44], passive spoofing via reflective surfaces [23], and other environmental exploits such as shadow/smoke [23, 62]. Beyond point-wise noise, *object-level* physical attacks seek adversarial 3D structures that persist across viewpoints and distances [8, 57, 64], and recent work emphasizes reproducible evaluation and sim-to-real transfer [10].

A recurring challenge in LiDAR attacks is the gap between simulation-time perturbations and real sensors, where beam patterns, range-dependent noise, and occlusion can substantially change point returns. This motivates both stronger physical modeling and robustness-oriented evaluation protocols, especially when the attacker must remain effective under realistic ego-motion and viewpoint changes.

Fusion Attacks and Object-level Adversaries. Multi-sensor fusion is designed to suppress modality-specific noise via cross-modal cues, making unified attacks harder than single-sensor ones. However, existing fusion attacks are still limited: many rely on unrealistic white-box access or perturb only one branch, while black-box attempts (e.g., frustum attack [15]) mainly target specific semantic-level fusion assumptions. Meanwhile, recent evidence [19, 40] shows that even single-modality perturbations can spill over to fusion detectors, implying that the fusion stack exposes non-trivial attack surfaces beyond individual sensors. From a system viewpoint, failures often arise from joint decision logic—misalignment between modalities, confidence domination of one branch, or error amplification in association/NMS—yet most attacks do not model or optimize against these coupling mechanisms. Finally, differentiable-rendering 3D adversaries [1, 33, 42, 43, 47, 54, 58] are predominantly texture-centric over predefined meshes, typically tied to digital rendering, and rarely validated for LiDAR or for cross-system robustness under realistic spatial variations.

These gaps motivate attacks that are black-box, object-level, and explicitly account for fusion coupling rather than optimizing a single modality in isolation. **Neural Rendering and Generative 3D Adversaries.** With the rise of neural scene/object representations, recent work [25, 29, 32, 61] explores adversarial objects via NeRF and 3D Gaussian Splatting (3DGS), enabling efficient multi-view optimization and rapid rendering. In parallel, generative 3D pipelines (text-to-3D, text-to-multi-view, and multi-view personalization) provide a natural mechanism to explore semantic variations beyond continuous texture noise, offering higher-level controllability and potentially better physical plausibility. Such controllable generative editing has also been systematically benchmarked for fine-grained attribute manipulation (e.g., drag-based gaze direction transformation) [13]. However, many existing neural-rendering attacks still rely on differentiable objectives and continuous parameter tuning, which may not match strict query-only settings and can overfit to limited viewpoints.

In contrast, we study a practical *object-level* threat model under strict *black-box* constraints, and search a discrete, text-conditioned generation space to induce meaningful geometry and appearance variations. Guided solely by confidence feedback from monocular and Camera–LiDAR fusion detectors, our approach yields physically plausible adversarial objects that transfer across perception pipelines and remain stable under diverse viewpoints and realistic driving conditions.

3 Method

Overview. Fig. 2 illustrates the overall three-stage workflow of Agent-OBJ. Starting from a base prompt p^{base} , an LLM agent iteratively edits the semantic/style prompts (p^{sem}, p^{sty}) to instantiate a candidate pedestrian-like 3D object (Sec. 3.2). The object is then rendered and inserted into calibrated camera–LiDAR observations via $\mathcal{R}$ and evaluated by a monocular 2D detector f_{2D} and a fusion 3D detector f_F over viewpoints $\mathcal{V}$ (Sec. 3.1). The resulting confidence feedback, together with naturalness and multi-view stability regularizers, defines a joint objective that guides prompt updates under a fixed query budget B (Sec. 3.3).

3.1 Problem Definition and Threat Model

We study a practical black-box, object-level attack on an autonomous-driving perception stack comprising a monocular 2D detector f_{2D} and a Camera–LiDAR fusion 3D detector f_F . The adversary physically inserts a single pedestrian-like object $\mathcal{O}$ at a fixed location in front of the ego vehicle and seeks targeted disappearance: $\mathcal{O}$ should be missed by both detectors over a set of realistic viewpoints $v \in \mathcal{V}$, instantiated by rotating the object’s yaw relative to the ego vehicle. The attacker is query-only: each query exposes only post-processed detections (bounding boxes and confidence scores), with no access to logits, gradients, intermediate features, or training data, and is constrained by a query budget B .

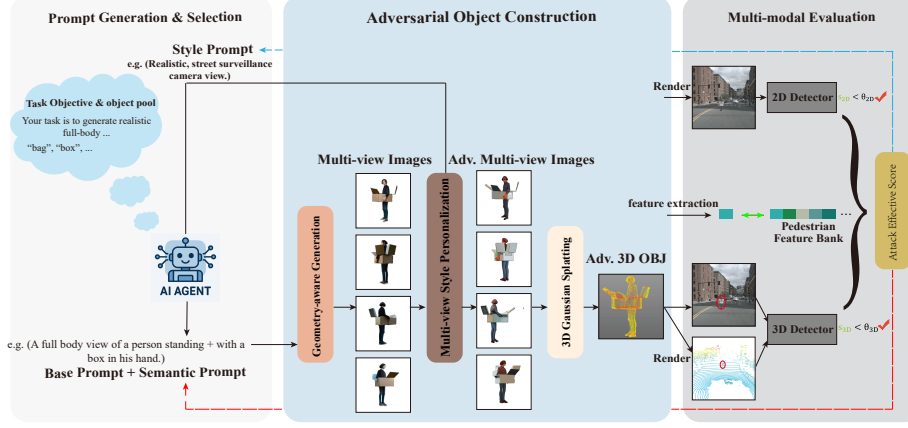


Fig. 2: The architecture of Agent-OBJ. This three-stage pipeline leverages an AI Agent for prompt steering, 3D Gaussian Splatting for adversarial object construction, and a multi-modal evaluation suite to evaluate disappearance attacks against state-of-the-art detectors.

Given a scene $\mathcal{S}$ and viewpoint v , rendering/insertion yields synchronized multi-sensor observations $(I_v, P_v) = \mathcal{R}(\mathcal{S}, \mathcal{O}, v)$, where I_v denotes the rendered camera images and P_v denotes the rendered LiDAR point cloud.

Insertion and rendering operator. We implement $\mathcal{R}$ as a calibration-consistent insertion operator that produces both camera and LiDAR observations. For cameras, $\mathcal{R}$ renders $\mathcal{O}$ into the target view using known extrinsics/intrinsics and alpha-composites it onto the background image while preserving depth ordering. For LiDAR, $\mathcal{R}$ places $\mathcal{O}$ in the ego frame and simulates its point returns using a beam-based forward model with range-dependent noise; the simulated returns are merged into the original scene point cloud. This design ensures each query reflects the detectors’ *final* post-processed predictions under a unified multi-sensor observation.

For detector $m \in \{2D, F\}$, we define the target confidence at v as the maximum target-class confidence among detections overlapping the object’s support $\mathcal{B}_v^\mathcal{O}$:

$$s_m(v) = \max_{b \in \mathcal{D}_m(\cdot)} \{ \text{conf}_{\text{target}}(b) \mid \text{IoU}(b, \mathcal{B}_v^\mathcal{O}) > \tau_{\text{IoU}} \}, \quad (1)$$

where $\mathcal{D}_m(\cdot)$ denotes the *final post-NMS* detections of f_m and we set $s_m(v) = 0$ if no box satisfies the overlap criterion.

Object support and matching criterion. The support $\mathcal{B}_v^\mathcal{O}$ is determined by the known insertion transform of $\mathcal{O}$. For f_{2D} , we obtain $\mathcal{B}_v^\mathcal{O}$ by projecting the inserted 3D bounding box onto the image plane and taking its tight 2D enclosing rectangle. For f_F , $\mathcal{B}_v^\mathcal{O}$ is the inserted 3D bounding box in the LiDAR/ego frame, and we

match predictions using *3D IoU* under the detector’s box parameterization. We apply an IoU gate τ_{IoU} to associate predicted boxes with the inserted object; if no post-NMS prediction overlaps $\mathcal{B}_v^{\mathcal{O}}$, we set $s_m(v) = 0$. The attack succeeds when both $s_{2D}(v)$ and $s_F(v)$ are suppressed below their respective decision thresholds.

Placement and viewpoints. We insert $\mathcal{O}$ at a fixed range along the ego vehicle’s forward axis so that it is visible in the front camera, and evaluate robustness under a discrete viewpoint set $\mathcal{V}$ realized by rotating the object’s yaw relative to the ego vehicle. In our default setting, $|\mathcal{V}| = 4$ (yaw angles specified in Sec. 4.1).

Decision thresholds and query accounting. We treat each detector as a black-box decision system with a fixed post-processing threshold θ_m (confidence and NMS). A viewpoint is considered *missed* if $s_m(v) < \theta_m$ or the overlapping detection is suppressed by NMS. Each evaluation of one object queries both detectors over all viewpoints, costing $2|\mathcal{V}|$ queries. Thus, T optimization iterations correspond to a total query budget of $B = 2T|\mathcal{V}|$ queries (or $T|\mathcal{V}|$ per detector).

Viewpoint-robust success criterion. We optimize for joint failures across $v \in \mathcal{V}$ rather than a single “lucky” pose. To quantify viewpoint robustness, we report the mean performance over $\mathcal{V}$ together with its cross-view variation, reflecting both effectiveness and stability.

3.2 Prompt-conditioned 3D Adversarial Object Generation

Decoupled prompts. We parameterize object generation with a base prompt p^{base} (e.g., “a full-body view of a person standing”), a *semantic* prompt p^{sem} that controls geometry/shape (e.g., “with a bag in hand”), and a *style* prompt p^{sty} that controls appearance (e.g., “realistic, street surveillance camera view”). This decoupling allows us to adjust shape and appearance with controllable degrees of freedom during optimization.

Why decoupling helps black-box optimization. In a strict query-only setting, directly optimizing continuous 3D parameters is prohibitively expensive. We instead search over a structured, text-conditioned space where edits to p^{sem} induce semantic/shape changes, while edits to p^{sty} primarily affect appearance. This factorization yields interpretable, low-entropy updates, improving query efficiency over unconstrained perturbations.

Multi-view synthesis and personalization. A text-to-multi-view generator produces view-consistent images for geometry initialization:

$$\mathcal{I}_{\text{geo}} = G_{\text{T2MV}}([p^{\text{base}}; p^{\text{sem}}]), \quad (2)$$

where $\mathcal{I}_{\text{geo}}$ contains a set of multi-view images with consistent structure. We then apply a pretrained multi-view personalization module, conditioned on multi-view images and p^{sty} , to obtain view-consistent adversarial appearance:

$$\mathcal{I}_{\text{adv}} = \Psi_{\text{edit}}(\mathcal{I}_{\text{geo}}, p^{\text{sty}}), \quad (3)$$

which edits texture/color/material while preserving multi-view consistency.

Two-stage multi-view consistency. We enforce view consistency twice: (i) G_{T2MV} stabilizes geometry by producing coherent multi-view structure, and (ii) Ψ_{edit} applies appearance edits conditioned on the multi-view set to prevent view-specific textures that break under yaw changes. This is critical because we evaluate and regularize performance over multiple viewpoints.

3D instantiation via 3DGS. Finally, we reconstruct a 3D Gaussian Splatting object from $\mathcal{I}_{adv}$:

$$\mathcal{O} = \text{3DGS}(\mathcal{I}_{adv}), \quad (4)$$

yielding a renderable object that supports photorealistic camera rendering and LiDAR simulation within $\mathcal{R}$.

3DGS representation. 3DGS represents $\mathcal{O}$ as a set of anisotropic Gaussians with parameters $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i(\cdot)\}_{i=1}^N$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ denote center and covariance, α_i is opacity, and $\mathbf{c}_i$ denotes view-dependent radiance. This representation supports fast differentiable rendering for cameras and efficient point sampling for LiDAR simulation, enabling repeated evaluation within the query budget.

Representation and controllable factors. In our prompt-conditioned pipeline, p^{sem} primarily drives *shape/structure* (e.g., body silhouette and carried items) through G_{T2MV} , while p^{sty} mainly affects *appearance* (e.g., clothing texture, color, material) through Ψ_{edit} . This separation enables targeted edits during optimization: semantic edits explore geometry-level variations, whereas style edits refine appearance while maintaining multi-view coherence.

3.3 Joint Black-box Optimization with Stealthiness and Multi-view Stability

Threshold-aware attack objective. Directly minimizing raw confidence can be suboptimal because detectors make decisions by thresholding. We therefore adopt a threshold-aware surrogate:

$$\phi_m(s) = \max(s - \theta_m, 0), \quad (5)$$

which penalizes confidence only when it exceeds the detector’s decision threshold θ_m . Our per-view attack loss is then

$$\begin{aligned} \mathcal{L}_{atk}(\mathcal{O}; v) &= \lambda_{2D} \phi_{2D}(s_{2D}(v)) + \lambda_F \phi_F(s_F(v)), \\ \mathcal{L}_{atk}(\mathcal{O}) &= \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathcal{L}_{atk}(\mathcal{O}; v). \end{aligned} \quad (6)$$

Joint aggregation across modalities. We optimize a unified objective that couples the two detectors at the score level rather than attacking each branch independently. In practice, we set $(\lambda_{2D}, \lambda_F)$ to balance the two confidence scales and prevent one branch from dominating optimization. This joint formulation is essential for achieving high ASR_{joint} , as single-modality objectives tend to overfit to one detector and transfer poorly to the other.

Stealthiness and stability regularization. We optimize a full objective that balances attack strength, naturalness, and cross-view robustness:

$$\mathcal{L}(\mathcal{O}) = \mathcal{L}_{atk}(\mathcal{O}) + \alpha \mathcal{L}_{nat}(\mathbf{X}) + \beta \mathcal{L}_{mv}(\mathcal{O}), \quad (7)$$

where $\mathbf{X}$ denotes the object’s geometry (point samples from 3DGS).

Naturalness regularizer $\mathcal{L}_{nat}$. To avoid unrealistic artifacts, we regularize the object to remain pedestrian-like in both geometry and semantics:

$$\mathcal{L}_{nat}(\mathbf{X}) = \underbrace{\|\mathbf{g}(\mathbf{X}) - \mathbf{g}_{ref}\|_1}_{\text{box-scale prior}} + \eta \underbrace{\text{Dist}(\Phi(\mathbf{X}), \mathcal{M}_{ped})}_{\text{feature plausibility}}, \quad (8)$$

where $\mathbf{g}(\mathbf{X})$ denotes the object’s 3D box size statistics (w, h, l) computed from the inserted object, and $\mathbf{g}_{ref}$ is the reference box-size statistic averaged over benign pedestrian instances. $\Phi(\mathbf{X})$ extracts a PointNet [34] embedding from the object point cloud, and $\mathcal{M}_{ped}$ is a pedestrian feature bank built by encoding nuScenes pedestrian point-cloud instances with the same PointNet. $\text{Dist}(\cdot, \mathcal{M}_{ped})$ is implemented as the minimum cosine distance to the bank.

Reference priors and feature bank. We use $\mathbf{g}(\mathbf{X}) = (w, h, l)$ as the box-scale prior and compute $\mathbf{g}_{ref}$ as the mean box dimensions from benign pedestrian instances. To model semantic plausibility, we construct $\mathcal{M}_{ped}$ by extracting PointNet embeddings from nuScenes pedestrian point-cloud instances under the same preprocessing and encoder. We define Dist as the minimum cosine distance to $\mathcal{M}_{ped}$, which penalizes out-of-distribution artifacts and aligns optimization with the stealthiness metric used in experiments.

Multi-view stability regularizer $\mathcal{L}_{mv}$. To avoid view-specific attacks, we encourage consistent effectiveness across viewpoints by penalizing cross-view variation:

$$\mathcal{L}_{mv}(\mathcal{O}) = \text{Std}_{v \in \mathcal{V}}[\mathcal{L}_{atk}(\mathcal{O}; v)], \quad (9)$$

which promotes stable suppression across $v \in \mathcal{V}$ rather than overfitting to a single pose.

Budget-aware update schedule. We adopt a simple schedule to allocate the limited budget across semantic and style edits. When the object becomes implausible (low naturalness), we prioritize semantic edits to restore structure, whereas when one detector remains confident, we adjust the complementary factor (shape or appearance) to reduce its score. This scheduling reduces wasted queries compared to unguided random edits.

Prompt-driven agent search under black-box feedback. We perform gradient-free optimization by iteratively editing (p^{sem}, p^{sty}) with an LLM agent under strict black-box feedback. The agent operates in a constrained action space of discrete prompt edits, including *add/replace/remove* of attribute phrases, while keeping

Algorithm 1: Agent-OBJ: Black-box prompt-driven optimization

Input: Scene $\mathcal{S}$, base prompt p^{base} , initial prompts (p_0^{sem}, p_0^{sty}) , viewpoints $\mathcal{V}$, budget B

Output: Adversarial object $\mathcal{O}^*$

Initialize $\mathcal{O}^* \leftarrow 3DGS(\Psi_{edit}(G_{T2MV}([p^{base}; p_0^{sem}]), p_0^{sty}))$;

Compute $\mathcal{L}(\mathcal{O}^*)$ by querying f_{2D}, f_F over all $v \in \mathcal{V}$;

Set number of iterations $T \leftarrow \lfloor B/(2|\mathcal{V}|) \rfloor - 1$ (the -1 accounts for the initial evaluation of $\mathcal{O}^*$);

for $t = 1$ **to** T **do**

- $(p_t^{sem}, p_t^{sty}) \leftarrow \text{EditAgent}(\mathcal{H}_{t-1})$;
- Instantiate $\mathcal{O}_t$ via Sec. 3.2;
- Query both detectors over all $v \in \mathcal{V}$ and compute $\mathcal{L}(\mathcal{O}_t)$;
- if** $\mathcal{L}(\mathcal{O}_t) < \mathcal{L}(\mathcal{O}^*)$ **then**
- Accept** edit: $\mathcal{O}^* \leftarrow \mathcal{O}_t$;
- else**
- Reject** edit: roll back to the previous prompts;
- Update history $\mathcal{H}_t$;
- Stop if budget B is exhausted;

return $\mathcal{O}^*$;

p^{base} fixed to preserve the pedestrian identity. At iteration t , the agent proposes one edit to either p^{sem} or p^{sty} based on a compact history $\mathcal{H}_{t-1}$ that summarizes recent prompts and scalar feedback $\{s_{2D}(v), s_F(v)\}_{v \in \mathcal{V}}$ along with regularizer values. We instantiate $\mathcal{O}_t$ via Sec. 3.2, query both detectors over all $v \in \mathcal{V}$, and compute $\mathcal{L}(\mathcal{O}_t)$. We maintain the best-so-far object $\mathcal{O}^*$; if $\mathcal{L}(\mathcal{O}_t) < \mathcal{L}(\mathcal{O}^*)$, we accept the update and set $\mathcal{O}^* \leftarrow \mathcal{O}_t$. Otherwise, we can keep the previous prompts (rollback) while recording the attempt for future edits. The procedure repeats until the query budget B is exhausted, and we report $\mathcal{O}^*$. Algorithm 1 summarizes the overall procedure.

Complexity and budget. Each iteration evaluates one instantiated object by querying both detectors over all viewpoints, requiring $2|\mathcal{V}|$ black-box queries. Therefore, the total number of queries scales linearly with iterations, i.e., $B = 2T|\mathcal{V}|$. The computational overhead outside detector queries is dominated by prompt-conditioned 3D instantiation (multi-view synthesis, personalization, and 3DGS reconstruction), which can be amortized by caching intermediate outputs when only p^{sty} changes while p^{sem} remains fixed.

Cost breakdown and caching. Each iteration incurs $2|\mathcal{V}|$ detector queries; thus the query complexity is $\mathcal{O}(T|\mathcal{V}|)$. We cache intermediate results whenever possible: if only p^{sty} changes, we reuse the geometry initialization from $G_{T2MV}([p^{base}; p^{sem}])$ and apply appearance edits on top, avoiding redundant generation. This amortization is beneficial under tight budgets where many iterations involve small appearance refinements.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on nuScenes [4], which contains 1,000 driving scenes (about 20 seconds each), around 40k annotated keyframes, roughly 1.4M camera images, and about 390k LiDAR sweeps. We use the official calibration and timestamps to generate synchronized multi-sensor observations.

Attack protocol. We select 10 scenes from the validation split and sample 5 keyframes per scene. For each keyframe, we insert a single pedestrian-like object at a fixed range (5–15 m) along the ego vehicle’s forward axis to ensure visibility in the front camera. This range covers the safety-critical zone where camera and LiDAR detection confidence peaks and the autonomous vehicle’s reaction constraints are most stringent; beyond 15 m, LiDAR sparsity already degrades all detectors independently of any attack. We evaluate robustness under four viewpoints $\mathcal{V}$ by rotating the object’s yaw relative to the ego vehicle while keeping its position fixed. Unless otherwise stated, results are averaged over all scenes, keyframes, and viewpoints.

Detectors. We consider a monocular 2D detector f_{2D} (YOLOv10x [46]) and a Camera–LiDAR fusion 3D detector f_F (MambaFusion [48]). We use official checkpoints with default inference settings and standard post-processing (NMS). During optimization, we use Qwen2.5 [35] as the language agent to propose prompt edits, and query only the detectors’ final outputs (bounding boxes and confidence scores) as black-box feedback.

Query budget. We optimize one adversarial object *per keyframe*. Each iteration evaluates one candidate by querying both detectors over all viewpoints in $\mathcal{V}$, costing $2|\mathcal{V}|$ queries. We run $T = 100$ iterations, resulting in a per-keyframe budget of $B = 2T|\mathcal{V}| = 800$ total queries (i.e., $4T$ per detector for $|\mathcal{V}| = 4$), and report the best-so-far object within the budget. Results are averaged over all keyframes.

4.2 Evaluation Metrics

We report Attack Success Rate (ASR, %) as the proportion of trials in which the inserted target is missed. Specifically, ASR_{2D} and ASR_F are computed w.r.t. the monocular and fusion detector, respectively, and ASR_{joint} requires both detectors to miss the target. We additionally report **confidence drop** as the reduction in the maximum target-class confidence before and after attack, denoted by Δs_{2D} and Δs_F . To quantify **stealthiness**, we measure (i) deviation of geometry statistics from benign references and (ii) feature distance to natural pedestrians (lower is better). Finally, we evaluate **multi-view stability** over $\mathcal{V}$ by reporting the mean ASR across viewpoints together with its cross-view variation (Std/Var), or worst-view ASR when applicable.

Implementation details. We instantiate a 3DGS object by reconstructing from four view-consistent images using a pretrained LGM [45]. The intermediate multi-view images are appearance-personalized with MV-Adapter [21] conditioned on

the style prompt p^{sty} , producing view-consistent texture/color/material edits at azimuths $0^\circ, 90^\circ, 180^\circ$, and 270° ; low-opacity Gaussians are filtered out. We evaluate robustness on the same four canonical viewpoints $\mathcal{V} = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, implemented as object-yaw rotations in the ego frame. We use an IoU gate $\tau_{IoU} = 0.5$ for association (2D projection for f_{2D} and 3D IoU in the LiDAR/ego frame for f_F). LiDAR returns are simulated via beam-based ray casting with distance-dependent range noise $\sigma(r) = \sigma_0 + kr$ using a 32-beam pattern, with $\sigma_0 = 0.02$ and $k = 10^{-4}$ (in meters), and are merged into the scene point cloud for MMDetection3D inference. We build the pedestrian feature bank $\mathcal{M}_{ped}$ by encoding 71,073 nuScenes pedestrian instances with PointNet [34] and define Dist as the minimum cosine distance to the bank. A query corresponds to running one detector at one viewpoint, hence the total budget is $B = 2T|\mathcal{V}|$. Unless stated otherwise, we use $\lambda_{2D} = \lambda_F = 1$, $\alpha = \beta = 0.1$, and $\eta = 1$. The EditAgent draws *add/replace/remove* operations from two disjoint pools, $\mathcal{D}_{sem}$ (geometry) and $\mathcal{D}_{sty}$ (texture), each containing ~ 30 candidate attribute phrases sampled independently, and retains only loss-reducing edits. We will release the full prompts, edit lists, and code to ensure reproducibility.

4.3 Main Results

As shown in Tab. 1, we report the main results on nuScenes. The Benign baseline inserts a normally generated pedestrian object without adversarial optimization. RandEdit removes agent guidance and instead applies random prompt edits from a predefined edit set, generating a different candidate object at each iteration under the same query budget. 2D-only and Fusion-only optimize the object using feedback from only the monocular detector or only the fusion detector, respectively.

Agent-OBJ achieves strong effectiveness on both branches ($ASR_{2D}=68.0$ and $ASR_F=96.0$), leading to a substantially higher joint success rate ($ASR_{joint}=64.0$) than all baselines. Compared with RandEdit and single-modality objectives, Agent-OBJ better balances unified attackability with robustness: it attains the lowest cross-view variation ($MV\text{-}Std=5.3$) while keeping FeatureDist close to the benign reference (0.12 vs. 0.08), indicating improved stability and perceptual plausibility under the same protocol.

4.4 Comparison with Prior Attacks

Tab. 2 reports representative prior attacks as *reference baselines*. These methods are *not directly comparable* to Agent-OBJ, since some assume white-box access and/or different threat models; inapplicable entries are marked as “—”. Image-space attacks (e.g., NI-FGSM) apply only to the camera branch, so we report ASR_{2D} only, while LiDAR/3D attacks (e.g., AdvPC) are reported in ASR_F only. Existing physical/object-level baselines (Invisible Attack and ScAR) show limited *simultaneous* failures, whereas Agent-OBJ achieves substantially higher unified impact ($ASR_{joint} = 64.0$) under a strict black-box, query-only setting.

Table 1: Main results on nuScenes (10 val scenes). We report attack success rates (ASR), confidence drop (Δs ; measured relative to the Benign Prompt-to-3D baseline), multi-view stability (Std over $|\mathcal{V}| = 4$ viewpoints), and stealthiness. “ $\pm$ ” denotes standard deviation over keyframes.

Method	Attack Effectiveness $\uparrow$			Confidence Drop $\uparrow$		Stealth & Stability $\downarrow$	
	ASR _{2D}	ASR _F	ASR _{joint}	Δs_{2D}	Δs_F	FeatureDist	MV-Std
Benign (Prompt-to-3D)	5.0 $\pm$ 2.1	20.0 $\pm$ 1.6	1.0 $\pm$ 0.7	–	–	0.08	2.7
RandEdit (query-matched)	14.1 $\pm$ 4.3	35.0 $\pm$ 8.7	13.9 $\pm$ 1.2	0.16	0.28	0.35	22.5
2D-only objective	78.0 $\pm$ 8.1	23.0 $\pm$ 3.0	20.3 $\pm$ 5.9	0.82	0.19	0.17	18.3
Fusion-only objective	15.0 $\pm$ 6.2	87.0 $\pm$ 3.5	13.7 $\pm$ 4.8	0.44	0.34	0.26	14.5
Agent-OBJ	68.0 $\pm$ 8.3	96.0 $\pm$ 3.7	64.0 $\pm$ 6.2	0.62	0.41	0.12	5.3

Table 2: Comparison with representative prior attacks. Image-only attacks (NI-FGSM, T-SEA, FCA, PGA) cannot report ASR_F/ASR_{joint}, and LiDAR-only attacks (AdvPC, ScAR) report ASR_F only; “–” marks inapplicable entries. Several baselines require white-box access or operate in a different threat model, thus are not directly comparable to our strict black-box setting. Agent-OBJ attains the highest ASR_{joint} despite the black-box, no-gradient constraint.

Method	Threat Model	ASR _{2D}	ASR _F	ASR _{joint}
NI-FGSM [27]	White-box, Img-Det	71.4	–	–
T-SEA [17]	White-box, Img-Det	57.2	–	–
FCA [47]	White-box, Img-Det	58.4	–	–
PGA [29]	White-box, Img-Det	61.3	–	–
AdvPC [16]	White-box, LiDAR	–	67.3	–
ScAR [30]	White-box, LiDAR	12.5	49.8	16.3
Invisible Attack [6]	White-box, Fusion	18.6	44.2	20.9
Frustum Attack [15]	Black-box, Fusion	–	87.3	26.1
Agent-OBJ (ours)	Black-box, Fusion	68.0	96.0	64.0

4.5 Ablation Studies

Tab. 3 ablates major components of Agent-OBJ and highlights their complementary roles. Removing the joint objective substantially reduces ASR_{joint} and increases cross-view variation, indicating that explicitly coupling the two detectors is critical for consistent unified failures. Disabling personalization further degrades ASR_{joint} to 38.4 and increases FeatureDist, showing that view-consistent appearance control is important for both effectiveness and plausibility.

The naturalness term $\mathcal{L}_{nat}$ reveals a clear effectiveness–stealth trade-off: dropping it boosts ASR_{joint} but significantly harms stealthiness and stability, with FeatureDist worsening and MV-Std increasing. Finally, removing the multi-view stability regularizer $\mathcal{L}_{mv}$ leads to the lowest ASR_{joint} (30.5) and higher variance, confirming that enforcing cross-view consistency is necessary to avoid viewpoint-specific “lucky” attacks.

Table 3: Ablation studies on nuScenes. Each component is removed from Agent-OBJ.

Variant	ASR _{joint} ↑	FeatureDist ↓	MV-Std ↓
Full (Agent-OBJ)	64.0	0.12	5.3
w/o joint objective	40.8	0.11	9.4
w/o personalization	38.4	0.15	8.2
w/o $\mathcal{L}_{nat}$	80.4	0.36	16.8
w/o $\mathcal{L}_{mv}$	30.5	0.14	8.3

Table 4: Transferability of the adversarial object optimized on (YOLOv10x, MambaFusion) in a black-box manner, reported *without re-optimization*. The last row additionally transfers the object *zero-shot* to a different dataset (KITTI).

Test setting	ASR _{2D}	ASR _F	ASR _{joint}
In-domain: (YOLOv10x, MambaFusion)	68.0	96.0	64.0
2D transfer: (DINO, MambaFusion)	57.6	96.0	53.7
Fusion transfer: (YOLOv10x, BEVFusion)	68.0	98.7	66.2
Cross-dataset: → KITTI [14] (zero-shot)	62.6	90.8	58.9

4.6 Transferability Across Detectors

Tab. 4 evaluates transferability without re-optimization. When swapping the 2D detector (YOLOv10x→DINO [59]) while keeping the fusion model fixed, the adversarial object still achieves strong ASR_F (96.0) and non-trivial joint success (53.7), suggesting that the learned adversary is not overfitted to a single monocular model. When swapping the fusion detector (MambaFusion→BEVFusion [26]) with the same 2D detector, ASR_{joint} further improves to 66.2, indicating that the generated object generalizes across different fusion architectures. Beyond cross-model transfer, the object also generalizes *across datasets*: applied zero-shot to KITTI [14], which differs in sensor configuration and calibration, it retains 58.9 ASR_{joint}, confirming that the attack exploits cross-modal fusion coupling rather than dataset-specific artifacts.

4.7 Query Efficiency

As shown in Tab. 5, Agent-OBJ achieves the best query efficiency under the same black-box protocol. RandEdit applies unguided random edits and does not use the returned scores to steer subsequent updates, so many queries are spent on low-yield directions, resulting in slower gains and lower ASR_{joint}. Greedy is more efficient by selecting the locally best score-improving edit, but its myopic nature can get trapped in local optima and saturate earlier as the budget grows. In contrast, Agent-OBJ uses score feedback to guide more informed exploration and higher-quality edits, making each query more likely to translate into progress; this is consistent with its lower Queries@80%, indicating faster convergence under a fixed budget.

Table 5: Query efficiency on nuScenes. ASR_{joint} under different query budgets (total detector calls over both detectors and all $|\mathcal{V}|$ viewpoints). Queries@80%: queries to reach 80% of the final ASR_{joint} at 800 queries.

Method	$ASR_{joint}@200$	$ASR_{joint}@400$	$ASR_{joint}@800$	Queries@80% ↓
RandEdit	2.8	6.7	13.9	712
Greedy	31.5	37.0	61.2	648
Agent-OBJ	37.0	45.9	64.0	584

5 Conclusion

Agent-OBJ exposes a critical blind spot in autonomous driving perception: a single, physically plausible 3D object can simultaneously degrade monocular and Camera–LiDAR fusion detectors, undermining the common redundancy-based safety assumption. At its core, Agent-OBJ employs an LLM agent to iteratively steer semantic and style prompts, searching a discrete text-conditioned generation space using only black-box confidence feedback from both detectors. Coupled with naturalness regularization and multi-view stability constraints, this agent-driven optimization produces stealthy and stable adversaries. Results on the nuScenes dataset highlight the urgency of developing stronger defenses against object-level, cross-modal attacks.

References

1. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: ICML (2018)
2. Bayer, J., Becker, S., Münch, D., Arens, M., Beyerer, J.: Higher-order adversarial patches for real-time object detectors. arXiv preprint arXiv:2601.04991 (2026)
3. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
5. Cao, Y., Bhupathiraju, S.H., Naghavi, P., Sugawara, T., Mao, Z.M., Rampazzi, S.: You can’t see me: Physical removal attacks on LiDAR-based autonomous vehicles driving frameworks. In: USENIX Security (2023)
6. Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., Chen, Q.A., Liu, M., Li, B.: Invisible for both camera and LiDAR: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In: S&P (2021)
7. Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q.A., Fu, K., Mao, Z.M.: Adversarial sensor attack on LiDAR-based perception in autonomous driving. In: ACM MM (2019)
8. Cao, Y., Xiao, C., Yang, D., Fang, J., Yang, R., Liu, M., Li, B.: Adversarial objects against LiDAR-based autonomous driving systems. arXiv preprint arXiv:1907.05418 (2019)

9. Cheng, J., Zhang, Y., Wang, Z., Qin, Z., Li, C.: DePatch: Towards robust adversarial patch for evading person detectors in the real world. arXiv preprint arXiv:2408.06625 (2024)
10. Cheng, L., Zhang, H., Zhang, L., Hermanns, H.: Revisiting physically realizable adversarial object attack against LiDAR-based detection: Clarifying problem formulation and experimental protocols. arXiv preprint arXiv:2507.18457 (2025)
11. Doan, B.G., Xue, M., Ma, S., Abbasnejad, E., Ranasinghe, D.C.: TNT attacks! universal naturalistic adversarial patches against deep neural network systems. *TIFS* **17**, 3816–3830 (2022)
12. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: *CVPR* (2018)
13. Fu, Y., Ding, J., Wang, R., Fu, Q., Tsang, I.W., Cheng, M.M., Guo, Q.: Benchmarking drag* for eye direction transformation and beyond. *Visual Intelligence* **3**(1), 29 (2025)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *CVPR* (2012)
15. Hallyburton, R.S., Liu, Y., Cao, Y., Mao, Z.M., Pajic, M.: Security analysis of Camera-LiDAR fusion against Black-Box attacks on autonomous vehicles. In: *USENIX Security* (2022)
16. Hamdi, A., Rojas, S., Thabet, A., Ghanem, B.: AdvPC: Transferable adversarial perturbations on 3D point clouds. In: *ECCV* (2020)
17. Huang, H., Chen, Z., Chen, H., Wang, Y., Zhang, K.: T-SEA: Transfer-based self-ensemble attack on object detection. In: *CVPR* (2023)
18. Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y.: Multi-modal sensor fusion for auto driving perception: A survey. arXiv preprint arXiv:2202.02703 (2022)
19. Huang, W., Han, A., Chen, Y., Cao, Y., Xu, Z., Suzuki, T.: On the comparison between multi-modal and single-modal contrastive learning. *NeurIPS* (2024)
20. Huang, Y., Ren, Y., Wang, J., Huo, L., Bai, X., Zhang, J., Yu, H.: AdvReal: Physical adversarial patch generation framework for security evaluation of object detection systems. *ESWA* **296**, 128967 (2026)
21. Huang, Z., Guo, Y.C., Wang, H., Yi, R., Ma, L., Cao, Y.P., Sheng, L.: MV-Adapter: Multi-view consistent image generation made easy. In: *ICCV* (2025)
22. Jin, Z., Ji, X., Cheng, Y., Yang, B., Yan, C., Xu, W.: PLA-LiDAR: Physical laser attacks against LiDAR-based 3D object detection in autonomous vehicle. In: *S&P* (2023)
23. Kobayashi, R., Nomoto, K., Tanaka, Y., Tsuruoka, G., Mori, T.: Invisible but detected: Physical adversarial shadow attack and defense on LiDAR object detection. In: *USENIX Security* (2025)
24. Lee, K., Chen, Z., Yan, X., Urtasun, R., Yumer, E.: ShapeAdv: Generating shape-aware adversarial 3D point clouds. arXiv preprint arXiv:2005.11626 (2020)
25. Li, L., Lian, Q., Chen, Y.C.: Adv3D: Generating 3D adversarial examples for 3D object detection in driving scenarios with NeRF. In: *IROS* (2024)
26. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: BEVFusion: A simple and robust LiDAR-camera fusion framework. *NeurIPS* (2022)
27. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. *ICLR* (2020)
28. Liu, H., Wu, Y., Yu, Z., Vorobeychik, Y., Zhang, N.: SlowLiDAR: Increasing the latency of LiDAR-based detection using adversarial examples. In: *CVPR* (2023)

29. Lou, T., Jia, X., Liang, S., Liang, J., Zhang, M., Xiao, Y., Cao, X.: 3D Gaussian splatting driven multi-view robust physical adversarial camouflage generation. In: ICCV (2025)
30. Lu, X., Radha, H.: ScAR: Scaling adversarial robustness for LiDAR object detection. In: IROS (2023)
31. Matos, F., Bernardino, J., Durães, J., Cunha, J.: A survey on sensor failures in autonomous vehicles: Challenges and solutions. *Sensors* **24**(16), 5108 (2024)
32. Meng, L., Shao, M., Qiao, Y., Liu, W., Lv, X.: Gaussian splitting attack: Gaussian splatting-based multi-view 3D adversarial attack. *Pattern Recognition* p. 112466 (2025)
33. Oslund, S., Washington, C., So, A., Chen, T., Ji, H.: Multiview robust adversarial stickers for arbitrary objects in the physical world. *JCCE* **1**(4), 152–158 (2022)
34. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: CVPR (2017)
35. Qwen Team: Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115 (2025)
36. Sato, T., Hayakawa, Y., Suzuki, R., Shiiki, Y., Yoshioka, K., Chen, Q.A.: LiDAR spoofing meets the new-gen: Capability improvements, broken assumptions, and new attack strategies. *NDSS* (2024)
37. Shi, H., Liu, W., Li, Z., Fang, X., Meng, X., Peng, W., Zhong, H., Liu, M., Wang, Y.: Intelligent robot systems: a survey from the perspective of visual intelligence. *Visual Intelligence* **4**, 14 (2026)
38. Shin, H., Kim, D., Kwon, Y., Kim, Y.: Illusion and dazzle: Adversarial optical channel exploits against LiDARs for automotive applications. In: CHES (2017)
39. Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., Kohno, T.: Physical adversarial examples for object detectors. In: WOOT (2018)
40. Su, Z., Xu, L., Zheng, Z., Yu, T., Liu, Y., Fang, L.: RobustFusion: Human volumetric capture with data-driven visual cues using an RGB-D camera. In: ECCV (2020)
41. Sun, J., Cao, Y., Chen, Q.A., Mao, Z.M.: Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In: USENIX Security (2020)
42. Suryanto, N., Kim, Y., Kang, H., Larasati, H.T., Yun, Y., Le, T.T.H., Yang, H., Oh, S.Y., Kim, H.: DTA: Physical camouflage attacks using differentiable transformation network. In: CVPR (2022)
43. Suryanto, N., Kim, Y., Larasati, H.T., Kang, H., Le, T.T.H., Hong, Y., Yang, H., Oh, S.Y., Kim, H.: ACTIVE: Towards highly transferable 3D physical camouflage for universal and robust vehicle evasion. In: ICCV (2023)
44. Suzuki, R., Sato, T., Hayakawa, Y., Ikeda, K., Sako, O., Nagata, R., Chen, Q.A., Yoshioka, K.: WIP: Towards practical LiDAR spoofing attack against vehicles driving at cruising speeds. In: VehicleSec (2024)
45. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: LGM: Large multi-view Gaussian model for high-resolution 3D content creation. In: ECCV (2024)
46. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: YOLOv10: Real-time end-to-end object detection. *NeurIPS* (2024)
47. Wang, D., Jiang, T., Sun, J., Zhou, W., Gong, Z., Zhang, X., Yao, W., Chen, X.: FCA: Learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack. In: AAAI (2022)
48. Wang, H., Gao, J., Hu, W., Zhang, Z.: MambaFusion: Height-fidelity dense global fusion for multi-modal 3D object detection. *ICCV* (2025)
49. Wang, J., Li, F., Zhang, X., Sun, H.: Adversarial obstacle generation against LiDAR-based 3D object detection. *TMM* **26**, 2686–2699 (2023)

50. Wang, X., Cai, M., Sohel, F., Sang, N., Chang, Z.: Adversarial point cloud perturbations against 3D object detection in autonomous driving systems. *Neurocomputing* **466**, 27–36 (2021)
51. Wei, X., Guo, Y., Yu, J., Zhang, B.: Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks. *TPAMI* **45**(7), 9041–9054 (2022)
52. Wei, X., Huang, Y., Sun, Y., Yu, J.: Unified adversarial patch for visible-infrared cross-modal attacks in the physical world. *TPAMI* **46**(4), 2348–2363 (2023)
53. Wei, X., Liang, S., Chen, N., Cao, X.: Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641* (2018)
54. Xiao, C., Yang, D., Li, B., Deng, J., Liu, M.: MeshAdv: Adversarial meshes for visual recognition. In: *CVPR* (2019)
55. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: *ICCV* (2017)
56. Xu, R., Liu, B., Zhang, K., Chen, H., Tao, D., Liu, W.: Noise-robust few-shot classification via variational adversarial data augmentation. *Computational Visual Media* **11**(1), 227–239 (2025)
57. Yang, K., Tsai, T., Yu, H., Panoff, M., Ho, T.Y., Jin, Y.: Robust roadside physical adversarial attack against deep learning in LiDAR perception modules. In: *ACM ASIACCS* (2021)
58. Zeng, X., Liu, C., Wang, Y.S., Qiu, W., Xie, L., Tai, Y.W., Tang, C.K., Yuille, A.L.: Adversarial attacks beyond the image space. In: *CVPR* (2019)
59. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605* (2022)
60. Zhang, S., Cheng, Y., Zhu, W., Ji, X., Xu, W.: CAPatch: Physical adversarial patch against image captioning systems. In: *USENIX Security* (2023)
61. Zhang, Y., Wang, L., Zhao, J., Zhao, W., Zhou, F., Dang, Y., Yin, J.: 3DGAA: Realistic and robust 3D Gaussian-based adversarial attack for autonomous driving. *arXiv preprint arXiv:2507.09993* (2025)
62. Zheng, S., Liu, W., Guo, Y., Zang, Y., Shen, S., Wang, C.: A new adversarial perspective for LiDAR-based 3D object detection. In: *AAAI* (2025)
63. Zhu, W., Ji, X., Cheng, Y., Zhang, S., Xu, W.: TPatch: A triggered physical adversarial patch. In: *USENIX Security* (2023)
64. Zhu, Y., Miao, C., Zheng, T., Hajiaghajani, F., Su, L., Qiao, C.: Can we use arbitrary objects to attack LiDAR perception in autonomous driving? In: *ACM CCS* (2021)