RASP: RESPONSE-AWARE SPATIAL AND POSE MANIPULATION ATTACKS FOR COLLABORATIVE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative perception improves the sensing capability of individual agents through feature sharing, but malicious feature messages can also introduce false detections, missed detections, and localization errors. Existing attacks often allocate perturbations using coarse regional labels, without accounting for whether selected responses remain distinct after non-maximum suppression or whether the poses of retained objects can be redirected under detection and shape constraints. To address these limitations, we propose RASP, a response-aware framework for spatial and pose manipulation attacks against collaborative perception. RASP consists of Spatial Response Allocation (SRA) and Constrained Pose Redirection (CPR). SRA uses bounded response probes to prioritize candidate locations under a fixed spatial budget, applies NMS-aware greedy selection to diversify spoofed detections, and suppresses residual responses around removal targets. CPR fixes dense output correspondences and redirects selected high-confidence objects toward prescribed positions and orientations, while enforcing a generation-stage score floor, regularizing unintended shape changes, and bounding the feature perturbation. Experiments on OPV2V and V2XSet show that RASP consistently reduces detection performance under no defense, ROBOSAC, and FLD, demonstrating effective manipulation attacks of collaborative predictions across different datasets and defense settings. The code and model will be publicly available.

1 INTRODUCTION

Collaborative perception enables connected autonomous vehicles to share complementary observations, extending sensing coverage beyond the ego vehicle's field of view (Chen et al., 2019; Hu et al., 2024a;b; Shi et al., 2020). Intermediate collaboration exchanges spatial features at lower communication cost than raw point clouds(Chen et al., 2022; Qiu et al., 2021), while selective transmission can further reduce bandwidth(Liu et al., 2020; Xu et al., 2025). This sharing also exposes fused detections to malicious feature messages (Chakraborty et al., 2018). In response, consistency-based defenses, including ROBOSAC(Li et al., 2023), MADE(Zhao et al., 2024), and CAD(Zhang et al., 2024), inspect prediction agreement, reconstructed features, or occupancy geometry. With these checks in place, an attacker must alter consequential predictions without making the shared information conspicuously inconsistent with the ego vehicle's observations.

Existing attacks address this challenge only in part. (1) Attacks that broadly optimize feature or detection changes do not explicitly preserve agreement with ego perception; large discrepancies can make their outputs conspicuous to consistency-based defenses (Madry et al., 2017; Carlini & Wagner, 2017; Tu et al., 2021). (2) Region-guided attacks use confidence partitions(Lin et al., 2025), blind-area masks(Tao et al., 2026b), or vulnerability maps(Tao et al., 2026a) to locate attack opportunities. Such region selection can preserve partial agreement, but it neither identifies candidate cells that respond to bounded perturbations nor prevents pose changes in retained high-confidence detections. (3) Proposal-level attacks either create persistent boxes to increase NMS latency(Wang et al., 2026) or optimize box-level objectives based on overlap or IoU(Tao et al., 2026a; Zhang et al., 2024). However, they do not jointly consider response attainability and post-NMS survival when selecting spatial targets, so multiple high-response candidates may still collapse into a single retained detection and waste the available budget.

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.

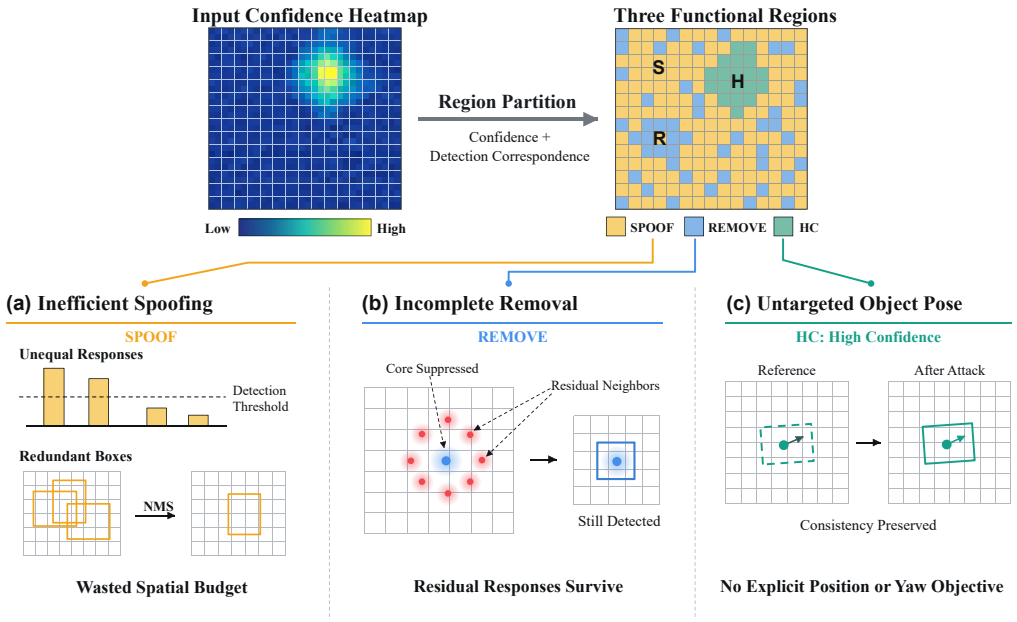


Figure 1: **Limitations of coarse regional objectives.** Within a SPOOF region, candidates can differ in achievable response and compete under NMS. Removing only the dominant response can leave nearby detections, while high-confidence (HC) consistency can coexist with pose errors.

Together, these gaps call for allocating a limited perturbation budget using the final fused detections. A spoofing candidate must cross the detection threshold, while nearby boxes may suppress one another under NMS and waste spatial slots. Removing a dominant detection can expose neighboring responses, and changing an existing object's pose requires enough confidence to keep it detectable. We therefore seek to coordinate response selection, neighborhood suppression, and constrained pose changes within a single perturbation. Fig. 1 illustrates these limitations of coarse regional objectives.

As illustrated in Fig. 2, we propose RASP, a response-aware attack with two components, namely Spatial Response Allocation (SRA) and Constrained Pose Redirection (CPR). SRA probes each SPOOF candidate at two bounded amplitudes and prioritizes locations predicted to cross the detection threshold. It then decodes their boxes and greedily selects a fixed number using pairwise BEV center distance and IoU as proxies for NMS compatibility. These geometric conditions are enforced when enough candidates are available; otherwise, a fallback fills the remaining slots. SRA also suppresses responses around REMOVE targets so that neighboring boxes do not replace the dominant detection. CPR fixes the output anchors of selected high-confidence objects and redirects their horizontal positions and orientations toward prescribed boxes. A score floor during optimization, shape regularization, and a feature amplitude bound encourage the redirected boxes to remain detectable while limiting unintended changes. Together, SRA and CPR control which responses receive the budget and how the geometry of retained objects changes under partial agreement with ego perception. Across OPV2V and V2XSet, RASP achieves the lowest defended AP among the evaluated attacks at IoU thresholds 0.3, 0.5, and 0.7. Its ROBOSAC/FLD bypass rates are $1.0/1.0$ on OPV2V and $0.97/0.95$ on V2XSet (Table 1).

Our main contributions are as follows:

- We propose a Response-Aware Spatial and Pose (RASP) manipulation attacks framework for collaborative perception.
- We develop Spatial Response Allocation (SRA), which leverage bounded response measurement, fixed-budget NMS-aware selection, and neighborhood suppression to target detections that can remain after post-processing.

- We design Constrained Pose Redirection (CPR) to steer selected high-confidence boxes toward prescribed positions and orientations while preserving partial agreement with ego perception.
- Extensive experiments on OPV2V and V2XSet demonstrate that RASP achieves state-of-the-art defended attack effectiveness while maintaining high defense-bypass rates.

2 RELATED WORK

2.1 ADVERSARIAL ATTACKS FOR COLLABORATIVE PERCEPTION

Attacks on collaborative perception have progressed from broad message perturbations(Tu et al., 2021) to targeted changes in feature fusion(Wang et al., 2025), confidence-defined regions(Lin et al., 2025), and ego blind areas(Tao et al., 2026b;a). Related work exploits detector post-processing(Wang et al., 2026) or subtly changes the poses of retained objects (Zhang et al., 2026). Together, these studies show that attack effectiveness depends on where a perturbation is applied and what remains in the final predictions. Yet region eligibility alone does not rank candidates by attainable response under a fixed budget, and promoting overlapping boxes can leave only one detection after NMS. Preserving a high-confidence detection also does not guarantee an accurate object pose. We investigate both limitations through response-aware spatial allocation and constrained pose redirection.

2.2 DEFENSES FOR COLLABORATIVE PERCEPTION

Defenses for collaborative perception assess shared messages through prediction consensus, feature consistency, or local geometry. ROBOSAC(Li et al., 2023) and CP-Guard(Hu et al., 2025) represent consensus-based approaches that compare collaborative outputs with ego perception. MADE(Zhao et al., 2024), FLD(Lin et al., 2025), and LUCIA(Wang et al., 2025) examine shared representations or adjust their influence at the feature level. CAD(Zhang et al., 2024) and PoseGuard(Zhang et al., 2026) add spatial checks using occupancy evidence or object-level anomalies. These defenses operate at different scales, but passing a consistency check does not guarantee the correctness of every detected object's score or pose. We examine whether bounded, localized feature changes can alter final detections while retaining enough agreement to evade such checks.

3 PRELIMINARIES

3.1 COLLABORATIVE PERCEPTION FRAMEWORK

We consider feature-level collaborative perception for 3D object detection, following the standard encoding–fusion–detection pipeline. Let $\mathcal{V} = \{0, \dots, N-1\}$ denote the agents participating in a frame, with agent 0 being the ego vehicle. Each agent i encodes its local LiDAR observation X_i into a bird's-eye-view (BEV) feature map. Writing $\mathcal{E}$ for the encoder and $\mathcal{T}_{i\to 0}$ for spatial alignment to the ego coordinate system, the features available for fusion are

$$f_i = \mathcal{E}(X_i), \qquad F_i = \mathcal{T}_{i\to 0}(f_i) \in \mathbb{R}^{C\times H\times W}, \tag{1}$$

where C is the number of channels and $H \times W$ is the BEV grid size.

The ego vehicle combines its own features with the received messages using a fusion module $\mathcal{A}$ and applies a detection head $\mathcal{H}$:

$$Z_0 = \mathcal{A}\big(F_0, \{F_i\}_{i\in\mathcal{V}\setminus\{0\}}\big), \qquad \widehat{Y}_0 = \mathcal{P}(\mathcal{H}(Z_0)), \tag{2}$$

where $\mathcal{P}$ denotes box decoding, score filtering, and non-maximum suppression. The prediction set $\widehat{Y}_0 = \{(b_k, s_k, c_k)\}_{k=1}^{K}$ contains 3D boxes, confidence scores, and class labels, respectively. Each box is parameterized by $b_k = (x_k, y_k, z_k, l_k, w_k, h_k, \psi_k)$, specifying its center, dimensions, and yaw. For an anchor-based head, we denote the classification logit and decoded box at BEV cell u and anchor a by $\ell_a(u)$ and $b_a(u)$. The corresponding vehicle score is $s_a(u) = \sigma(\ell_a(u))$, where σ is the sigmoid function.

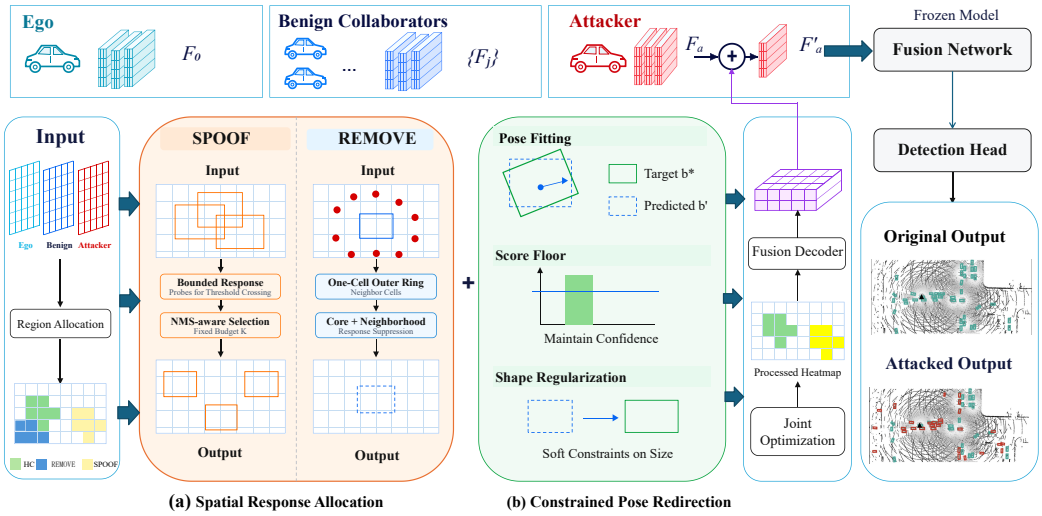


Figure 2: **Overview of the RASP framework.** Ego predictions define region-wise adversarial targets. SRA allocates spoofing responses and suppresses removal responses, while CPR redirects the position and yaw of selected HC objects. Joint optimization applies a bounded perturbation to the attacker's shared feature and uses the resulting detections as feedback.

3.2 Threat Model

Attack formulation. For offline white-box generation, the attacker knows the perception model, current-frame features, and reference annotations, and modifies only outgoing messages from a fixed collaborator subset $\mathcal{C} \subseteq \mathcal{V} \setminus \{0\}$. All other inputs remain fixed; defense access is not assumed.

Let $\widetilde{F}_i$ be the message supplied by compromised agent i, and define $F_i^{\mathrm{adv}} = \widetilde{F}_i$ for $i \in \mathcal{C}$ and $F_i^{\mathrm{adv}} = F_i$ otherwise. Let Φ denote the undefended ego perception pipeline in Eq. 2. The attacked output used for generation is

$$\widehat{Y}_0^{\mathrm{adv}} = \Phi\big(F_0, \{F_i^{\mathrm{adv}}\}_{i \in \mathcal{V}\setminus\{0\}}\big)\,. \tag{3}$$

The attack objective is

$$\max_{\widetilde{F}_{\mathcal{C}} \in \Omega_{\mathcal{C}}} \mathcal{J}\Big(\widehat{Y}_0^{\mathrm{adv}}, Y_{\mathrm{ref}}\Big)\,, \tag{4}$$

where $\widetilde{F}_{\mathcal{C}} = \{\widetilde{F}_i\}_{i \in \mathcal{C}}$, $\Omega_{\mathcal{C}}$ is the admissible message set, and $\mathcal{J}$ measures missed or fabricated detections and pose errors relative to Y_{ref}. Defense bypass is evaluated after generation.

4 Methodology

4.1 Framework Overview

As illustrated in Fig. 2, RASP follows four stages. (1) Current-frame features are used to obtain reference heatmaps and fused detections. (2) Ego confidence and fused detection correspondence partition targets into **SPOOF** for false detections, **REMOVE** for suppressed detections, and **HC** for pose changes to retained high-confidence objects. (3) Spatial Response Allocation (SRA) assigns a fixed SPOOF budget using bounded probes and NMS-aware selection and suppresses neighboring REMOVE responses; Constrained Pose Redirection (CPR) adjusts selected HC boxes. Their regional objectives are optimized jointly under a feature-amplitude bound. (4) The modified feature messages are sent to the ego vehicle for fusion and detection. Region assignments and selected spatial supports remain fixed during optimization.

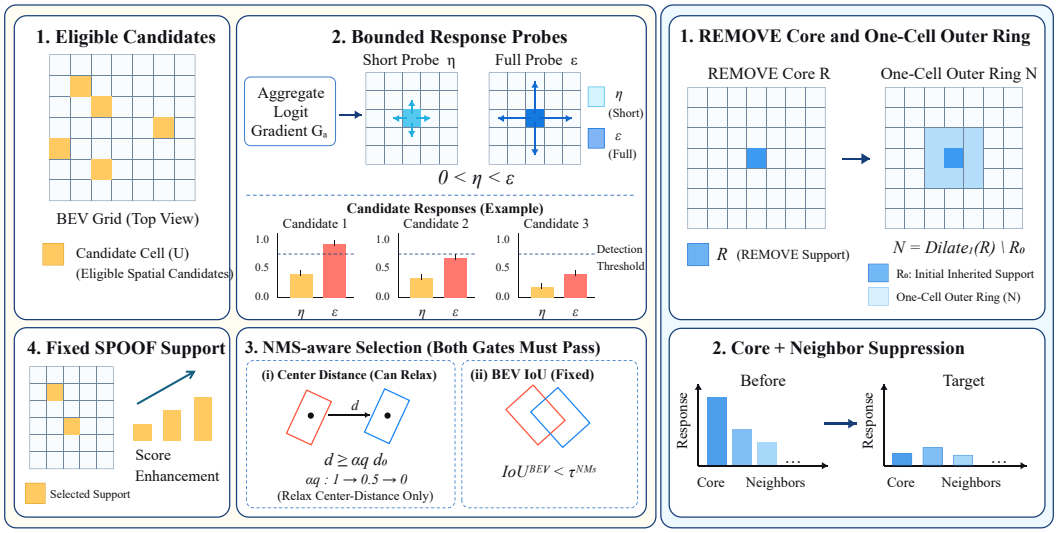


Figure 3: Spatial Response Allocation (SRA). (a) Bounded response probes rank SPOOF candidates, and NMS-aware selection allocates a fixed spatial budget for score enhancement. (b) REMOVE suppresses core detections and neighboring responses through a one-cell outer ring. The REMOVE neighborhood is excluded before SPOOF allocation.

4.2 Spatial Response Allocation

Fig. 3 illustrates SRA's two operations, namely allocating a fixed SPOOF budget and suppressing core and neighboring REMOVE responses. For controlled collaborator $i \in \mathcal{C}$, let $\bar{F}_i$ be its fixed reference message and $F_i' = \widetilde{F}_i = \bar{F}_i + \Delta_i$ its modified message. We write Δ for the joint perturbation and measure amplitude bounds relative to $\bar{F}_i$. At BEV cell u, the maximum vehicle score is $s'(u) = \max_a \sigma(\ell_a(u;\Delta))$, where a indexes anchors.

NMS-aware response allocation in SPOOF. Region labels alone do not identify SPOOF cells likely to cross the detection threshold under a bounded perturbation, and nearby boxes may compete under NMS. SRA constructs the candidate set $\mathcal{U}$ from geometrically eligible locations in the SPOOF region. For each anchor type a, we compute the aggregate response gradient at the clean feature and construct short- and full-amplitude probes:

$$G_a = \nabla_\Delta \sum_{u\in\mathcal{U}} \ell_a(u;\Delta)\Big|_{\Delta=0}, \qquad D_a^\rho = \rho M_\mathcal{U} \odot \operatorname{sign}(G_a), \quad \rho \in \{\eta, \epsilon\}, \tag{5}$$

where $0 < \eta < \epsilon$ and $M_\mathcal{U}$ restricts the probes to the candidate support. Candidates are ranked by their bounded probe responses, prioritizing likely threshold crossings.

A score-only top-K rule can spend several selections on neighboring cells whose boxes suppress one another during non-maximum suppression. To avoid this, we reuse each full probe's classification and regression outputs and associate candidate u with the decoded box b_u of its winning probe and output anchor. Starting from an empty set $\mathcal{S}$, the main selection pass scans candidates in response order and accepts them at ladder tier q only if

$$\|c(b_u) - c(b_v)\|_2 \geq \alpha_q d_0, \qquad \mathrm{IoU}_{\mathrm{BEV}}(b_u, b_v) < \tau_{\mathrm{nms}}, \quad \forall v \in \mathcal{S}, \tag{6}$$

where $c(\cdot)$ is the BEV center. We use $d_0 = 3.2$ m, $\tau_{\mathrm{nms}} = 0.15$, and $(\alpha_0, \alpha_1, \alpha_2) = (1, 0.5, 0)$. Candidates rejected at one tier remain available at the next, while the pairwise IoU constraint is retained throughout the ladder. Candidates whose decoded boxes overlap clean ground truth by at least 0.3 BEV IoU are deprioritized, favoring unoccupied regions. The selector uses a fixed budget of $K = 10$ cells, which share the SPOOF score-enhancement objective.

Neighborhood response suppression in REMOVE. After suppressing the confidence of the dominant detection box, neighboring dense candidate boxes may still be detected. SRA further suppresses the confidence of these residual detections through a one-cell-wide, eight-connected outer ring. Let $\mathcal{R}_0$ denote the initial inherited regional support and $\mathcal{R}$ the REMOVE support defined by M_R. The neighborhood $\mathcal{N}$ and its loss are defined as

$$\begin{aligned} \mathcal{N} &= \mathrm{Dilate}_1(\mathcal{R}) \setminus \mathcal{R}_0, \\ \mathcal{L}_N &= -\sum_{u \in \mathcal{N}} \log(1 - s'(u)) \,. \end{aligned} \tag{7}$$

The ring is constructed before SPOOF allocation, excluded from $\mathcal{U}$, and fixed throughout optimization.

For a reference REMOVE box b, let b' and s' denote the corresponding perturbed box and confidence score. Its core suppression loss is

$$\ell_R(b) = -\,\mathrm{IoU}(b', b) \log(1 - s') \,. \tag{8}$$

The core term acts on reference detections, whereas the neighborhood term suppresses nearby dense responses without requiring a reference object at every ring cell. Both belong to the REMOVE objective.

4.3 Constrained Pose Redirection

Preserving high-confidence detections does not ensure accurate object poses. CPR therefore steers the positions and orientations of selected HC boxes while constraining their detection scores and shapes.

Eligible objects have sufficiently high clean scores, anchor locations inside HC, and adequate observation support from a controlled collaborator. Each object is assigned to the collaborator providing the strongest support, measured through a voxel-based point-count proxy. For each selected object, a target box b^* specifies a horizontal displacement and yaw change while retaining the reference vertical position and dimensions.

Let b' and s' be the perturbed box and score at the same fixed anchor as b. To isolate pose changes, fitting uses $\check{b}' = (x', y', z, l, w, h, \psi')$, combining predicted horizontal position and yaw with reference vertical position and dimensions. We compute $I = \mathrm{sIoU}(\check{b}', b^*)$ using sigmoid-smoothed box occupancies on a fixed BEV grid. Writing $d = (z, l, w, h)$ and $d' = (z', l', w', h')$ for the reference and predicted shape-related components, the per-object objective is

$$\ell_H(b) = \lambda_p \log(\max\{1 - I, \zeta\}) + \lambda_f [\tau_f - s']_+ + \lambda_s \,\mathrm{SmoothL1}_\beta(d' - d), \tag{9}$$

where $[x]_+ = \max(x, 0)$, τ_f is the generation-stage score floor, and $\beta = 0.1$. For a vector argument, $\mathrm{SmoothL1}_\beta$ denotes the sum of the component-wise penalties. The score-floor term penalizes only scores below τ_f; the shape term softly discourages changes to the actual decoded box's vertical position and dimensions.

4.4 Region-wise Objective

We define one loss contribution for each functional region. Let $\mathcal{D}_H$ and $\mathcal{D}_R$ denote the selected HC objects and reference REMOVE detections, respectively, with $Q = |\mathcal{D}_H|$. Using the per-object losses $\ell_H(b)$ and $\ell_R(b)$ from Eq. 9 and 8, the regional objective is

$$\mathcal{L}_\nu = \begin{cases} \dfrac{1}{Q} \displaystyle\sum_{b \in \mathcal{D}_H} \ell_H(b), & \nu = H, \\ \displaystyle\sum_{b \in \mathcal{D}_R} \ell_R(b) + \lambda_N \mathcal{L}_N, & \nu = R, \\ -\displaystyle\sum_{u \in \mathcal{S}} \log(s'(u) + \zeta) \,, & \nu = S, \end{cases} \tag{10}$$

where H, R, and S refer to HC, REMOVE, and SPOOF, respectively. The joint optimization problem is

$$\min \mathcal{L}, \qquad \mathcal{L} = \mathcal{L}_H + \mathcal{L}_R + \mathcal{L}_S. \tag{11}$$

Table 1: Attack comparison on OPV2V and V2XSet using PointPillars with average fusion. The best AP results are shown in **bold**, and the second-best results are underlined.

Method	Dataset	No Defense AP			ROBOSAC AP			ASR	FLD AP			ASR
		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	
No Attack	OPV2V	0.91	0.91	0.81	0.91	0.91	0.81	-	0.91	0.91	0.81	-
PGD		0.63	0.59	0.18	0.83	0.82	0.67	0.0	0.91	0.91	0.81	0.0
C&W		0.42	0.40	**0.13**	0.85	0.83	0.69	0.0	0.91	0.91	0.81	0.0
James' Attack		0.58	0.55	0.19	0.85	0.83	0.69	0.0	0.91	0.91	0.81	0.0
AdvGPS		0.46	0.45	0.33	0.64	0.64	0.54	0.57	0.91	0.91	0.81	0.0
PB		0.52	0.51	0.45	0.51	0.50	0.44	1.0	0.52	0.51	0.45	1.0
RASP (Ours)		**0.36**	**0.35**	0.22	**0.36**	**0.35**	**0.22**	1.0	**0.36**	**0.35**	**0.22**	1.0
No Attack	V2XSet	0.84	0.83	0.73	0.84	0.83	0.73	-	0.84	0.83	0.73	-
PGD		0.77	0.74	0.42	0.84	0.83	0.72	0.0	0.84	0.83	0.73	0.0
C&W		0.63	0.59	0.29	0.84	0.83	0.72	0.0	0.84	0.83	0.73	0.0
James' Attack		0.67	0.63	0.30	0.84	0.83	0.72	0.0	0.84	0.83	0.73	0.0
AdvGPS		0.51	0.48	0.34	0.58	0.56	0.42	0.81	0.84	0.82	0.72	0.0
PB		0.51	0.47	0.35	0.51	0.48	0.36	0.98	0.51	0.48	0.36	0.95
RASP (Ours)		**0.44**	**0.33**	**0.17**	**0.45**	**0.34**	**0.18**	0.97	**0.44**	**0.35**	**0.18**	0.95

Table 2: Attack comparison on OPV2V with different detector and fusion combinations. The best AP results are shown in **bold**, and the second-best results are underlined.

Method	Detector+Fusion Method	No Defense AP			ROBOSAC AP			ASR	FLD AP			ASR
		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	
No Attack	PointPillars+Attention	0.95	0.94	0.83	0.95	0.94	0.83	-	0.95	0.94	0.83	-
PGD		0.21	0.14	**0.01**	0.89	0.87	0.71	0.03	0.95	0.94	0.83	0.0
C&W		0.24	0.16	0.02	0.86	0.82	0.61	0.11	0.87	0.85	0.70	0.04
James' Attack		**0.18**	**0.12**	**0.01**	0.86	0.80	**0.17**	0.59	0.88	0.87	0.73	0.03
PB		0.58	0.55	0.48	0.58	0.55	0.48	1.0	0.58	0.55	0.48	1.0
RASP (Ours)		0.39	0.36	0.25	**0.39**	**0.36**	0.25	1.0	**0.39**	**0.36**	**0.25**	1.0
No Attack	SECOND+Average	0.92	0.92	0.89	0.92	0.92	0.89	-	0.92	0.92	0.89	-
PGD		0.89	0.87	0.30	0.92	0.91	0.84	0.15	0.92	0.92	0.89	0.0
C&W		0.81	0.78	0.44	0.92	0.91	0.86	0.14	0.92	0.92	0.89	0.0
James' Attack		0.71	0.68	**0.12**	0.91	0.90	0.84	0.02	0.92	0.92	0.89	0.0
PB		0.68	0.64	0.45	0.68	0.64	0.45	1.0	0.68	0.64	0.45	1.0
RASP (Ours)		**0.40**	**0.36**	0.21	**0.40**	**0.36**	**0.21**	1.0	**0.40**	**0.36**	**0.21**	1.0

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and models. We evaluate on the validation sets of OPV2V (Xu et al., 2022b) and V2XSet (Xu et al., 2022a), with up to seven collaborating agents. The cross-dataset comparison uses PointPillars (Lang et al., 2019) with average fusion. On OPV2V, we additionally evaluate PointPillars with attentive fusion and SECOND with average fusion; the ablations use PointPillars with attentive fusion.

Attack settings. We set the perturbation budget to $\epsilon = 0.5$ and run each attack for 30 optimization iterations. For RASP, the NMS-aware selection threshold is $\tau_{\text{nms}} = 0.15$ and the initial center distance is $d_0 = 3.2\,\text{m}$. The SRA and CPR components are described in Section 4.

Baselines and metrics. We use PGD (Madry et al., 2017), C&W (Carlini & Wagner, 2017), James' Attack (Tu et al., 2021), AdvGPS (Li et al., 2024), and PB (Lin et al., 2025) as attack baselines. We evaluate detection performance using average precision (AP) at IoU thresholds of 0.3, 0.5, and 0.7 (Xu et al., 2022b); a larger AP drop from the no-attack result indicates a more effective attack. For defense evaluation, we use ROBOSAC (Li et al., 2023) and the feature-level

Table 3: Module ablation on OPV2V with PointPillars and attentive fusion. Lower AP indicates stronger attack effectiveness.

Method	No Defense AP			ROBOSAC AP			ASR	FLD AP			ASR
	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	
No Attack	0.95	0.94	0.83	0.95	0.94	0.83	-	0.95	0.94	0.83	-
RASP (Ours)	0.39	0.36	0.25	0.39	0.36	0.25	1.0	0.39	0.36	0.25	1.0
w/o SRA	0.53	0.47	0.33	0.53	0.47	0.33	1.0	0.53	0.47	0.33	1.0
w/o CPR	0.40	0.38	0.30	0.40	0.38	0.30	1.0	0.40	0.38	0.30	1.0

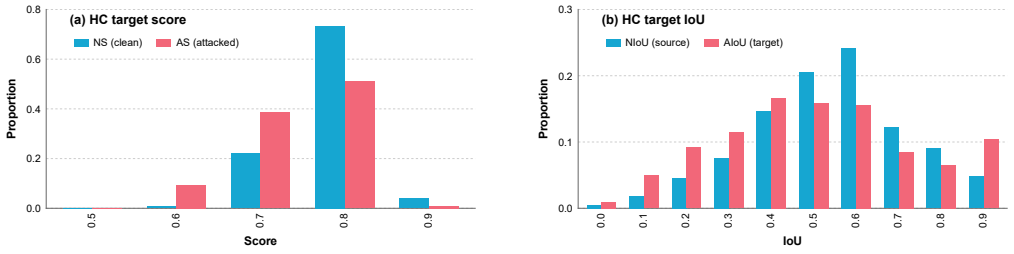


Figure 4: HC target distributions. (a) NS and AS denote clean and attacked scores. (b) NIoU and AIoU measure the corresponding predictions' IoU with ground truth.

defense FLD (Lin et al., 2025). The defense-bypass attack success rate (ASR) measures how often an attack evades a defense and serves as our stealthiness metric.

5.2 Main Results

Results on OPV2V and V2XSet. Table 1 compares attacks on both datasets. On OPV2V, RASP achieves AP@0.3/0.5/0.7 of $0.36/0.35/0.22$ under all three defense conditions, with an ASR of 1.0 against both defenses. Compared with PB, AP@0.7 decreases by 0.23 without defense and under FLD, and by 0.22 under ROBOSAC. On V2XSet, RASP lowers PB's AP@0.7 by 0.18 in each condition while maintaining ASRs of 0.97 and 0.95 under ROBOSAC and FLD, respectively. Although some baselines achieve lower undefended AP on OPV2V, RASP yields the lowest defended AP at all three thresholds on both datasets. These results show stronger detection degradation than PB with comparable defense-bypass rates.

Different detectors and fusion methods. Table 2 evaluates alternative configurations on OPV2V. With PointPillars and attentive fusion, RASP reduces AP@0.3/0.5/0.7 from PB's $0.58/0.55/0.48$ to $0.39/0.36/0.25$. With SECOND and average fusion, the corresponding values decrease from $0.68/0.64/0.45$ to $0.40/0.36/0.21$. These results hold under all three defense conditions, with ASR remaining 1.0. Thus, the improvement over PB extends across the evaluated detectors and fusion methods. James' Attack obtains lower AP@0.7 under ROBOSAC with attentive fusion (0.17), but its ASR is 0.59, compared with 1.0 for RASP.

5.3 Ablation Studies

Effect of SRA and CPR. We conduct ablations on OPV2V with PointPillars and attentive fusion. As shown in Table 3, removing SRA increases AP@0.3/0.5/0.7 from $0.39/0.36/0.25$ to $0.53/0.47/0.33$, while removing CPR increases the metrics to $0.40/0.38/0.30$. Both modules contribute to attack effectiveness. SRA has a larger impact at the lower IoU thresholds, whereas CPR's largest contribution occurs at IoU 0.7, consistent with its emphasis on object geometry. All variants retain an ASR of 1.0 under ROBOSAC and FLD.

Table 4: SRA component ablation on OPV2V with PointPillars and attentive fusion.

Method	No Defense AP			ROBOSAC AP			ASR	FLD AP			ASR
	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	
No Attack	0.95	0.94	0.83	0.95	0.94	0.83	-	0.95	0.94	0.83	-
RASP (Ours)	0.39	0.36	0.25	0.39	0.36	0.25	1.0	0.39	0.36	0.25	1.0
w/o NMS-aware selection	0.45	0.42	0.31	0.45	0.42	0.31	1.0	0.45	0.42	0.31	1.0
w/o neighborhood suppression	0.44	0.40	0.27	0.44	0.40	0.27	1.0	0.44	0.40	0.27	1.0

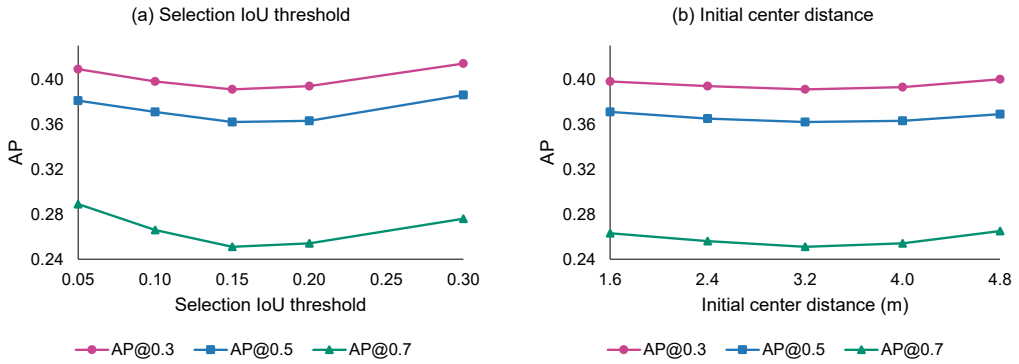


Figure 5: Sensitivity of NMS-aware selection on OPV2V with PointPillars and attentive fusion under no defense. (a) Varying τ_{nms} with $d_0 = 3.2$ m. (b) Varying d_0 with $\tau_{\text{nms}} = 0.15$. The SPOOF budget is fixed at $K = 10$. Lower AP indicates stronger attacks.

HC confidence and localization. Fig. 4 examines the effect on HC targets. Attacked scores remain concentrated around 0.7–0.8, while the IoU distribution changes, with less mass in the 0.5–0.8 bins. This shows that geometric changes can coexist with strong detection responses, consistent with CPR's aim of altering object poses while maintaining sufficient scores.

Effect of SRA components. Table 4 evaluates NMS-aware selection and neighborhood suppression. Removing NMS-aware selection raises AP@0.3/0.5/0.7 to $0.45/0.42/0.31$, while removing neighborhood suppression raises the metrics to $0.44/0.40/0.27$. Both changes weaken the attack without changing the reported ASR. These results support the use of spatially diversified spoofing responses and suppression of neighboring removal responses within SRA.

Sensitivity to selection geometry. We separately vary the selection IoU threshold τ_{nms} and the initial center distance d_0 on OPV2V with PointPillars and attentive fusion under no defense, while keeping all other settings, including the detector's NMS threshold, fixed. Fig. 5 shows that AP is lowest near the default settings and varies only modestly under nearby parameter choices. Within the tested ranges, AP is more sensitive to τ_{nms} than to d_0. This is consistent with our selection procedure: the IoU constraint is retained across all selection tiers, whereas the center-distance constraint is progressively relaxed.

6 Conclusion

In this paper, we presented RASP for stealthy attacks on collaborative perception. Spatial Response Allocation (SRA) assigns a fixed spoofing budget to NMS-aware locations and suppresses neighboring responses around removal targets, while Constrained Pose Redirection (CPR) shifts selected high-confidence boxes under score and shape constraints. Together, they induce false detections, missed detections, and pose errors while retaining plausible high-confidence responses. Experiments on OPV2V and V2XSet show the lowest defended AP among the evaluated attacks and high bypass rates under ROBOSAC and FLD, highlighting its effectiveness.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing by improving the grammar, wording, and readability of the manuscript and providing critical feedback on the clarity of the methodology description and experimental presentation. The research ideas, methodology, implementation, experiments, data, and reported results were developed and produced by the authors. We reviewed, verified, and revised all AI-assisted textual edits. We take responsibility for the final content of this work, including its technical claims and experimental results.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3.2 states the attacker's control and objective, while Section 4 specifies the SRA response probes, NMS-aware SPOOF selection, REMOVE neighborhood suppression, CPR constraints, and joint optimization objective. Section 5.1 reports the OPV2V and V2XSet validation sets, detector and fusion settings, perturbation and optimization parameters, baselines, defenses, and AP/ASR metrics. Section 5.3 tests the contributions of SRA and CPR; Appendices A–D provide target-attainment measurements, additional-defense results, qualitative BEV comparisons, and the SPOOF-budget scan. Furthermore, after this paper is accepted, we will publicly release the source code and models.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Hanlin Chen, Brian Liu, Xumiao Zhang, Feng Qian, Z Morley Mao, and Yiheng Feng. A cooperative perception environment for traffic operations and control. *arXiv preprint arXiv:2208.02792*, 2022.

Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on distributed computing systems (ICDCS)*, pp. 514–524. IEEE, 2019.

Senkang Hu, Zhengru Fang, Haonan An, Guowen Xu, Yuan Zhou, Xianhao Chen, and Yuguang Fang. Adaptive communications in collaborative perception with domain alignment for autonomous driving. In *GLOBECOM 2024-2024 IEEE Global Communications Conference*, pp. 746–751. IEEE, 2024a.

Senkang Hu, Zhengru Fang, Yiqin Deng, Xianhao Chen, Yuguang Fang, and Sam Kwong. Toward full-scene domain generalization in multi-agent collaborative bird's eye view segmentation for connected and autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 26 (2):1783–1796, 2024b.

Senkang Hu, Yihang Tao, Guowen Xu, Yiqin Deng, Xianhao Chen, Yuguang Fang, and Sam Kwong. Cp-guard: Malicious agent detection and defense in collaborative bird's eye view perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23203–23211, 2025.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12689–12697. IEEE, 2019.

Jinlong Li, Baolu Li, Xinyu Liu, Jianwu Fang, Felix Juefei-Xu, Qing Guo, and Hongkai Yu. Advgps: Adversarial gps for multi-agent perception attack. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18421–18427. IEEE, 2024.

Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among us: Adversarially robust collaborative perception by consensus. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 186–195. IEEE, 2023.

Hongwei Lin, Dongyu Pan, Qiming Xia, Hai Wu, Cheng Wang, Siqi Shen, and Chenglu Wen. Pretend benign: A stealthy adversarial attack by exploiting vulnerabilities in cooperative perception. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19947–19956. IEEE, 2025.

Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4105–4114. IEEE, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Hang Qiu, Pohan Huang, Namo Asavisanu, Xiaochen Liu, Konstantinos Psounis, and Ramesh Govindan. Autocast: Scalable infrastructure-less cooperative perception for distributed collaborative driving. *arXiv preprint arXiv:2112.14947*, 2021.

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10526–10535. IEEE, 2020.

Yihang Tao, Senkang Hu, Haonan An, Zhengru Fang, Hangcheng Cao, and Yuguang Fang. Learning mutual view information graph for adaptive adversarial collaborative perception. *arXiv preprint arXiv:2602.19596*, 2026a.

Yihang Tao, Senkang Hu, Yue Hu, Haonan An, Hangcheng Cao, and Yuguang Fang. Gcp: guarded collaborative perception with spatial-temporal aware malicious agent detection. *IEEE Transactions on Dependable and Secure Computing*, 2026b.

James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7748–7757. IEEE, 2021.

Chenyi Wang, Raymond Muller, Ruoyu Song, Jean-Philippe Monteuuis, Jonathan Petit, Yanmao Man, Ryan Gerdes, Z Berkay Celik, and Ming Li. From threat to trust: Exploiting attention mechanisms for attacks and defenses in cooperative perception. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 7387–7406, 2025.

Chenyi Wang, Ruoyu Song, Raymond Muller, Jean-Philippe Monteuuis, Z Berkay Celik, Jonathan Petit, Ryan Gerdes, and Ming Li. CP-FREEZER: Latency attacks against vehicular cooperative perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1114–1122, 2026.

Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6834–6843. IEEE, 2025.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pp. 107–124. Springer, 2022a.

Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589. IEEE, 2022b.

Qingzhao Zhang, Shuowei Jin, Ruiyang Zhu, Jiachen Sun, Xumiao Zhang, Qi Alfred Chen, and Z Morley Mao. On data fabrication in collaborative vehicular perception: Attacks and countermeasures. In *33rd USENIX security symposium (USENIX security 24)*, pp. 6309–6326, 2024.

Qingzhao Zhang, Runting Zhang, and Z Morley Mao. From stealthy data fabrication to unsafe driving: Realistic scenario attacks on collaborative perception. *arXiv preprint arXiv:2605.01301*, 2026.

Yangheng Zhao, Zhen Xiang, Sheng Yin, Xianghe Pang, Yanfeng Wang, and Siheng Chen. Made: Malicious agent detection for robust multi-agent collaborative perception. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13817–13823. IEEE, 2024.

A POSE-TARGET ATTAINMENT OF CPR

The target-fitting diagnostic complements the detector-level AP ablation in Table 3. Here IoU is computed against prescribed attack target boxes, rather than ground truth as in the distributions of Fig. 4. For each target, %Imp is the fraction for which the attacked prediction has higher target IoU than the clean prediction. The two %S columns report the fractions reaching target IoU thresholds of 0.5 and 0.7; IoU (Δ) gives the mean target IoU and its change from the clean prediction.

Table 5: CPR target-fitting metrics on OPV2V with PointPillars and attentive fusion. IoU is measured against the prescribed attack target; Δ is the gain over the clean prediction.

Method	%Imp	%S ($IoU_{0.5}$)	%S ($IoU_{0.7}$)	IoU (Δ)
No Attack	0.0	17.9	0.3	0.330 (0.000)
RASP (Ours)	96.4	66.8	35.4	0.603 (+0.273)
w/o CPR	21.1	8.0	0.1	0.298 (−0.032)

With CPR, RASP improves target IoU for 96.4% of the prescribed targets and reaches target IoU $0.5/0.7$ on $66.8\%/35.4\%$. Its mean target IoU is 0.603, compared with 0.330 for the clean prediction. Omitting CPR reduces these values to 21.1%, $8.0\%/0.1\%$, and 0.298, respectively. The diagnostic therefore supports CPR's role in following prescribed poses; the scene-level detection effect is assessed separately by the AP ablation.

B PERFORMANCE UNDER ADDITIONAL DEFENSES

Table 6 extends the OPV2V comparison to LUCIA, PoseGuard, and CP-Guard. It reports AP at three IoU thresholds and the defense-bypass ASR for each defense.

Table 6: OPV2V attack performance under no defense, LUCIA, PoseGuard, and CP-Guard: AP at IoU $0.3/0.5/0.7$ and defense-bypass ASR.

Method	No Defense AP			LUCIA AP			ASR	PoseGuard AP			ASR	CP-Guard AP			ASR
	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$		$IoU_{0.3}$	$IoU_{0.5}$	$IoU_{0.7}$	
No Attack	0.95	0.94	0.83	0.95	0.94	0.83	–	0.95	0.94	0.83	–	0.95	0.94	0.83	–
James' Attack	0.18	0.12	0.01	0.65	0.47	0.17	0.36	0.59	0.38	0.09	0.16	0.89	0.88	0.74	0.03
PB	0.58	0.55	0.48	0.58	0.55	0.48	1.0	0.58	0.55	0.48	1.0	0.58	0.55	0.48	1.0
RASP (Ours)	0.39	0.36	0.25	0.39	0.36	0.25	1.0	0.39	0.36	0.25	1.0	0.39	0.36	0.25	1.0

RASP has an ASR of 1.00 under each of the three tested defenses. Its AP remains 0.39, 0.36, and 0.25 at IoU thresholds 0.3, 0.5, and 0.7, respectively. PB also retains an ASR of 1.00, but its AP is higher at 0.58, 0.55, and 0.48. James' Attack has lower AP without defense, yet its ASR falls to 0.36, 0.16, and 0.03 under LUCIA, PoseGuard, and CP-Guard, respectively. Under CP-Guard, its AP rises to 0.89, 0.88, and 0.74. These results distinguish attack strength without defense from bypass of the defenses evaluated here.

C QUALITATIVE BEV COMPARISONS

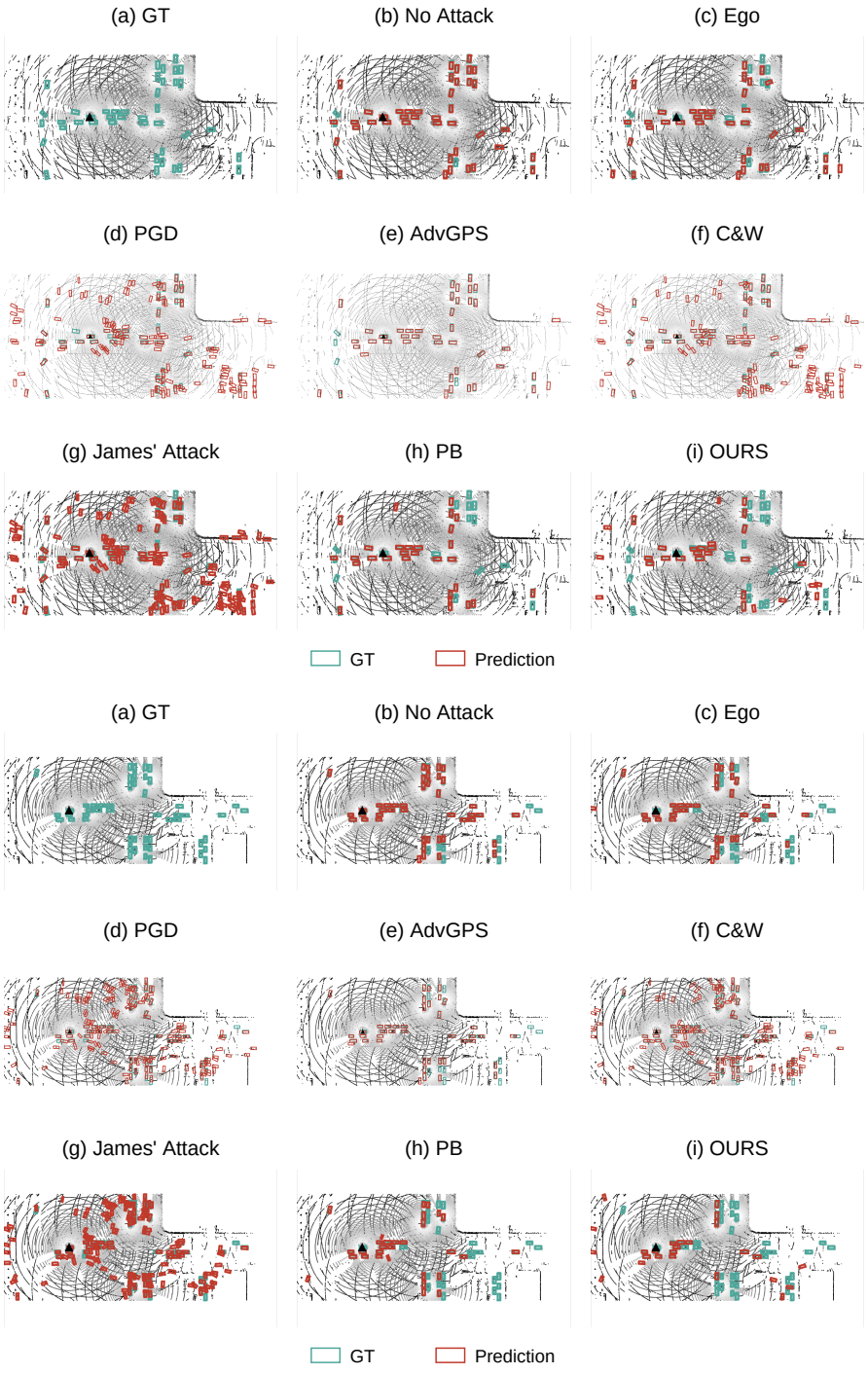


Figure 6: Visualization comparison of attack results between RASP and other attack methods.

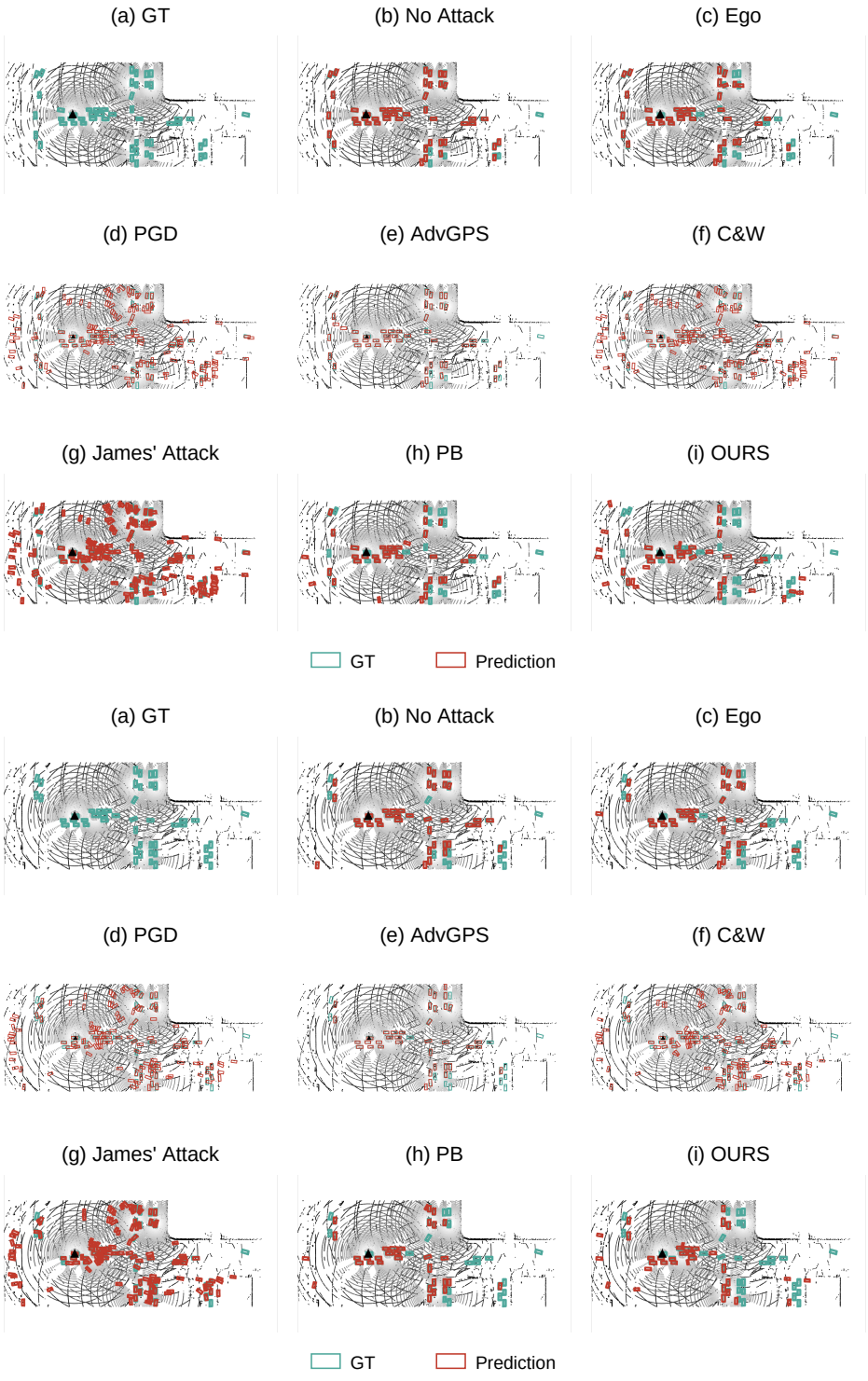


Figure 7: Visualization comparison of attack results between RASP and other attack methods.

Fig. 6 and 7 show four selected validation frames. In each nine-panel grid, the panels show ground truth, no attack, ego-only perception, PGD, AdvGPS, C&W, James' Attack, PB, and RASP in that order. In these examples, PGD, C&W, and James' Attack produce denser sets of red boxes away from ground-truth objects, whereas PB and RASP retain sparser prediction patterns. RASP also leaves some ground-truth boxes without nearby predictions or shifts the alignment of retained predictions. These examples illustrate the spatial character of the attacks; dataset-level comparisons are given by AP in Table 1.

D SPOOF BUDGET SENSITIVITY

Fig. 8 compares NMS-aware selection (ON) with naive top-K selection (OFF) as the SPOOF budget increases on OPV2V without defense. Lower AP indicates a stronger attack. At $K = 5$, ON and OFF achieve AP@0.5 of 0.4300 and 0.4480, respectively, a gap of only 0.0180. With five selected locations, there are fewer opportunities for naive top-K choices to compete under NMS; the small observed gap is consistent with this hypothesis.

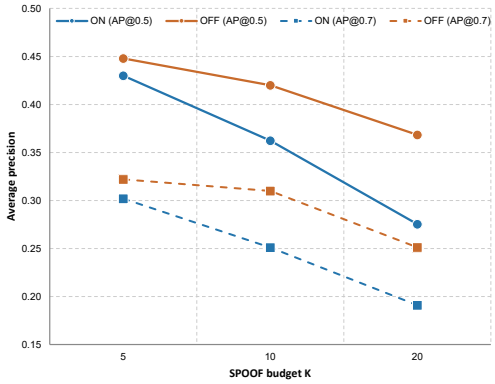


Figure 8: SPOOF-budget sensitivity of NMS-aware selection on OPV2V without defense. ON uses NMS-aware selection; OFF uses naive top-K. Solid curves show AP@0.5 and dashed curves show AP@0.7; lower AP indicates a stronger attack.

The gap widens with budget. At $K = 10$, AP@0.5 is 0.3624 for ON and 0.4200 for OFF (gap 0.0576); at $K = 20$, it is 0.2754 versus 0.3684 (gap 0.0930). From $K = 5$ to $K = 20$, AP@0.5 falls by 0.1546 with ON, compared with 0.0796 with OFF. The AP@0.7 curves show the same trend: the OFF–ON gap rises from 0.0200 at $K = 5$ to 0.0602 at $K = 20$.

This widening gap is consistent with naive top-K increasingly spending locations on decoded boxes that compete under NMS, while NMS-aware selection favors distinct effective targets. At $K = 20$, OFF reaches AP@0.5 of 0.3684, slightly above the 0.3624 attained by ON at $K = 10$, despite using twice as many locations. This scan tests budget sensitivity; the main comparison retains its fixed $K = 10$ budget.