Safety in Embodied AI: A Survey of Risks, Attacks, and Defenses

Xiao Li[1,*], Xiang Zheng[3,*], Yifeng Gao[1], Xinyu Xia[4], Yixu Wang[1], Xin Wang[1], Ye Sun[1], Yunhan Zhao[1], Ming Wen[1,2], Jiayu Li[1], Zixing Chen[1], Xun Gong[4], Yi Liu[3], Yige Li[5], Yutao Wu[6], Cong Wang[3], Ran He[13], Jun Sun[5], Yixin Cao[1,2], Zhineng Chen[1], Jingjing Chen[1], Tao Gui[1,2], Qi Zhang[1], Zuxuan Wu[1,2], Xipeng Qiu[1,2], Xuanjing Huang[1], Tiehua Zhang[7], Zhipeng Wei[9], Hanxun Huang[10], Sarah Erfani[10], James Bailey[10], Jianping Wang[3], Wei-Ying Ma[3,11], Chaowei Xiao[12], Bo Li[8], Xingjun Ma[1,2,†], Yu-Gang Jiang[1,†]

[1]Fudan University, [2]Shanghai Innovation Institute, [3]City University of Hong Kong
[4]Jilin University, [5]Singapore Management University, [6]Deakin University, [7]Tongji University, [8]UIUC
[9]UC Berkeley, [10]The University of Melbourne, [11]Tsinghua University, [12]Johns Hopkins University,
[13]Institute of Automation, Chinese Academy of Sciences

[*]Equal Contribution, [†]Corresponding authors

Abstract

Embodied Artificial Intelligence (Embodied AI) integrates perception, cognition, planning, and interaction into agents that operate in open-world, safety-critical environments. As these systems gain autonomy and enter domains such as transportation, healthcare, and industrial or assistive robotics, ensuring their safety becomes both technically challenging and socially indispensable. Unlike digital AI systems, embodied agents must act under uncertain sensing, incomplete knowledge, and dynamic human–robot interactions, where failures can directly lead to physical harm. This survey provides a comprehensive and structured review of safety research in embodied AI, examining attacks and defenses across the full embodied pipeline, from perception and cognition to planning, action & interaction, and agentic system. We introduce a multi-level taxonomy that unifies fragmented lines of work and connects embodied-specific safety findings with broader advances in vision, language, and multimodal foundation models. Our review synthesizes insights from over 400 papers spanning adversarial, backdoor, jailbreak, and hardware-level attacks; attack detection, safe training and robust inference; and risk-aware human–agent interaction. This analysis reveals several overlooked challenges, including the fragility of multimodal perception fusion, the instability of planning under jailbreak attacks, and the trustworthiness of human–agent interaction in open-ended scenarios. By organizing the field into a coherent framework and identifying critical research gaps, this survey provides a roadmap for building embodied agents that are not only capable and autonomous but also safe, robust, and reliable in real-world deployment.

Correspondence: xingjunma@fudan.edu.cn; ygj@fudan.edu.cn
Website: https://github.com/x-zheng16/Awesome-Embodied-AI-Safety
Key Words: Embodied AI Safety; Trustworthy Embodied AI; Multimodal Safety; Attacks and Defenses

[*]Equal Contribution.

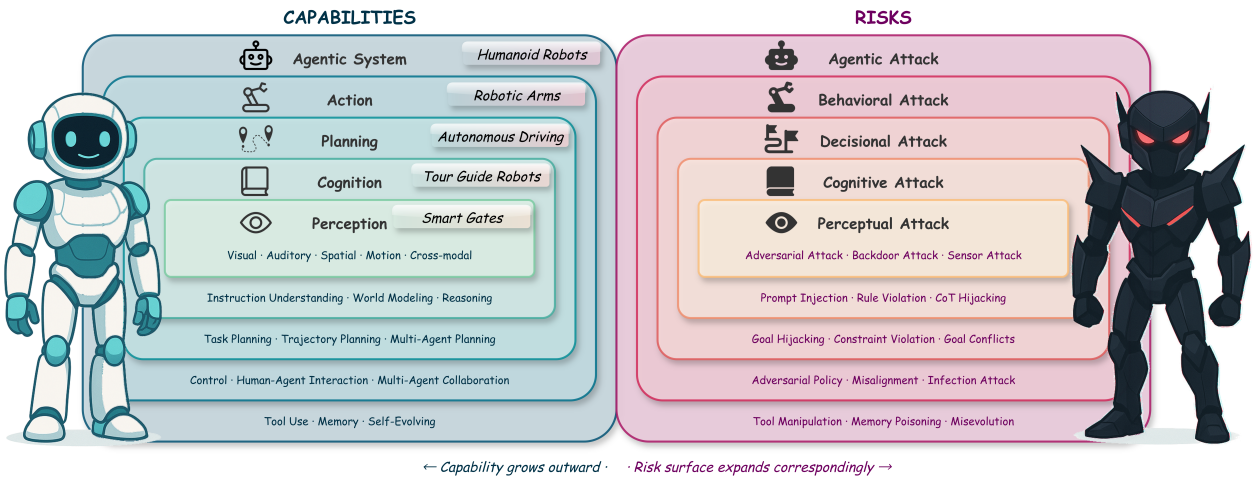

Figure 1 Capability vs. risk duality in embodied AI systems. **Left:** Nested capability layers from perception (innermost) to agentic systems (outermost), with representative embodiments at each level: sensor-only devices (e.g., face-recognition access controls), dialogue robots (e.g., museum guides), autonomous vehicles, robotic arms and humanoids, and future agentic robots with memory and tool use. **Right:** Corresponding safety risks at each layer. As capabilities expand outward, the attack surface grows correspondingly—vulnerabilities at inner layers cascade to outer layers, amplifying risks in more autonomous systems.

1 Introduction

Embodied Artificial Intelligence (Embodied AI) seeks to endow autonomous agents with the ability to perceive, reason, plan, and interact with the physical world [340]. Unlike purely digital AI systems, embodied agents operate in dynamic, uncertain, and safety-critical environments such as autonomous driving [253, 333, 479], collaborative robotics [30, 327], smart healthcare [98, 128, 167], and assistive robotics [11, 116]. In these settings, unsafe perception, flawed reasoning, erroneous planning, or unsafe interaction can lead not only to degraded task performance but also to real-world accidents, physical harm, and loss of human trust.

Recent years have witnessed rapid advances across the embodied AI pipeline [30, 176, 282, 286, 384, 385]. Improvements in perception (e.g., vision, LiDAR, multimodal sensing), cognition (e.g., world modeling, value alignment), planning (e.g., task planning, trajectory optimization), and interaction (e.g., safe control, human–robot collaboration) have significantly expanded the capabilities of embodied agents. However, these capabilities also broaden and complicate the attack surface [85, 365, 481]. Safety challenges that once appeared primarily in digital domains, such as adversarial examples in vision or jailbreak prompts in language models, carry far more severe consequences in physical environments. For instance, small perturbations to a visual sensor may cause an autonomous vehicle to misinterpret a stop sign [89], while maliciously poisoned training data may compromise task planning and produce unsafe trajectories [85]. Misaligned or unpredictable human–agent interactions can further generate behaviors that endanger users directly.

Despite their growing importance, the safety challenges unique to embodied AI remain underexamined. Existing surveys in AI safety largely focus on digital-only systems such as vision foundation models [251, 252, 358], large language models (LLMs), multimodal large language models (MLLMs) [417], or digital agents [75, 103]. While these works offer valuable taxonomies of attacks and defenses, they rarely address embodied settings where perception, cognition, planning, and interaction are tightly coupled and must operate under real-world constraints. A comprehensive treatment of embodied AI safety therefore requires not only synthesizing research within each component but also integrating insights from broader AI safety domains that have direct implications for embodied systems.

Capability–Risk Duality. A key organizing principle of this survey is the capability–risk duality: each layer

†Corresponding authors.

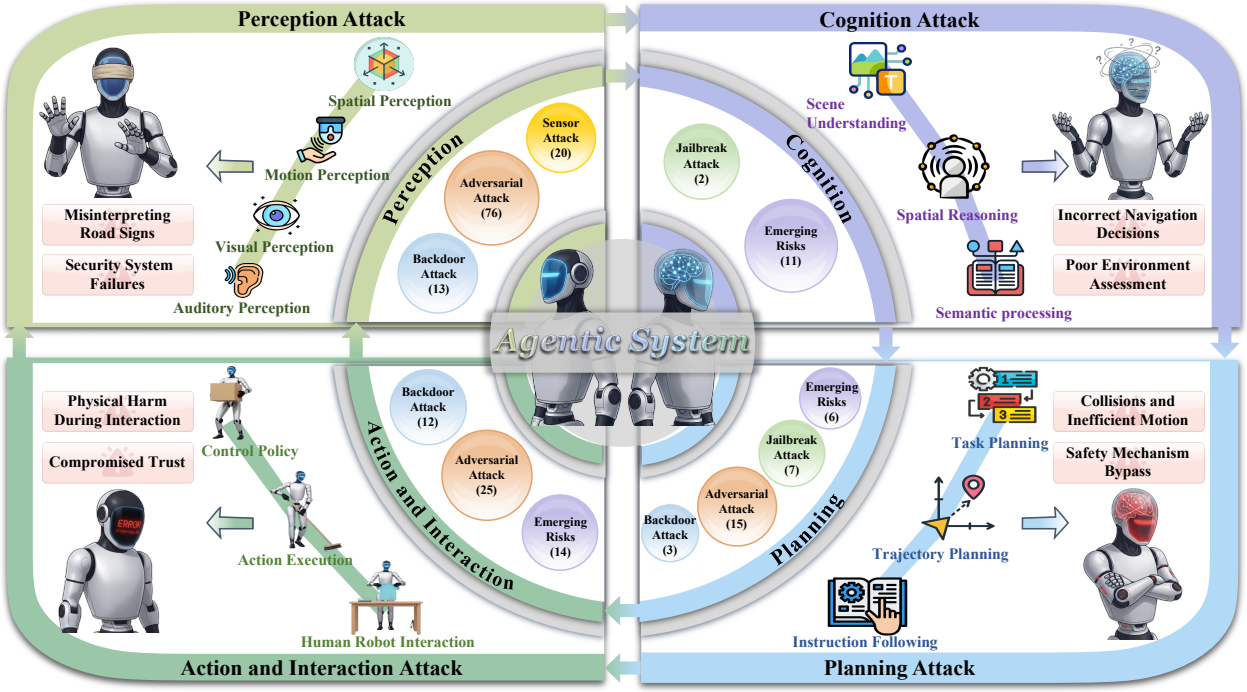

Figure 2 Illustration of safety threats and attack surfaces across capability layers of embodied AI systems.

of the embodied pipeline represents not merely a functional component but a capability expansion that introduces corresponding new vulnerabilities, as illustrated in Figure 1. Real-world embodied systems vary in the depth of this capability stack. At the innermost layer, sensor-only devices such as face-recognition access controls represent the simplest embodied systems, where adversaries can only target perceptual inputs. Adding cognition yields agents like museum guide robots capable of dialogue and scene understanding, opening attack surfaces in reasoning and language comprehension. Incorporating planning enables navigation and decision-making, as in autonomous vehicles, where adversaries can additionally manipulate route planning and trajectory prediction. At the action and interaction layer, robotic arms and humanoid robots gain the ability to physically manipulate their environment, exposing control and human–robot interaction to exploitation. Finally, agentic systems augment all prior capabilities with persistent memory, tool use, and self-evolution, creating the broadest attack surface where compromises at any inner layer can cascade outward. This duality (deeper capability entails broader risk) motivates our layered taxonomy and structures the remainder of this survey: each section addresses both the attacks specific to its capability layer and the pathways through which inner-layer vulnerabilities propagate to outer-layer failures.

To address this need, we conduct a systematic survey of **safety research in embodied AI**. We propose a multi-level taxonomy that organizes vulnerabilities and defenses across five key components of embodied AI systems: **perception**, **cognition**, **planning**, **action & interaction**, and **agentic system**. For each component, we categorize attacks and defenses, including adversarial perception, unsafe reasoning, planning under perturbations, unsafe control and interaction, and agentic-level risks such as tool misuse, memory poisoning, and cascading failures, as illustrated in Figure 2. We further extend our analysis to include highly relevant studies from traditional AI safety, spanning vision, language, and multimodal foundation models, with a focus on those that have clear implications for embodied intelligence. Figure 3 presents an overview of different attack and defense types and their distribution across the pipeline components. This dual perspective situates embodied AI safety within the broader AI safety ecosystem while highlighting the unique risks that emerge when intelligence is deployed in the physical world.

Based on the current literature, we identify and summarize the threats posed by various attacks, as shown in

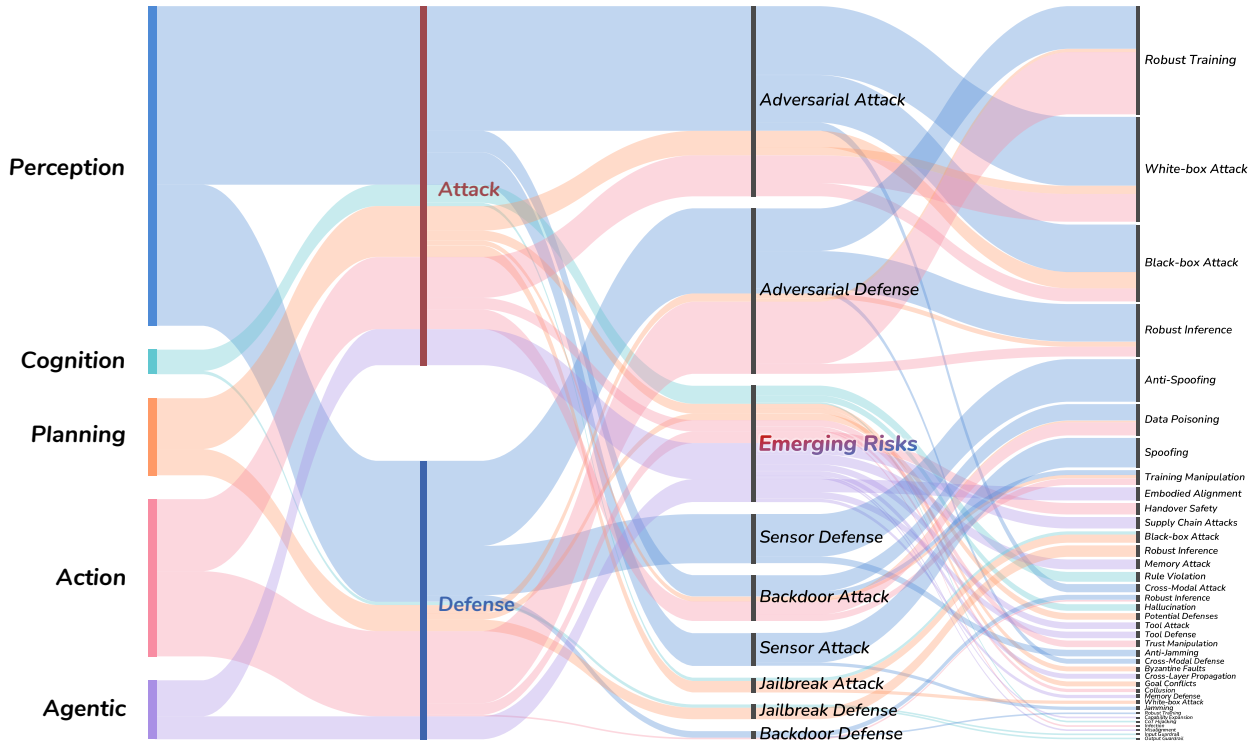

Figure 3 Overview of representative attack and defense methods across perception, cognition, planning, action & interaction, and agentic system layers. The width of the strips is proportional to the number of reviewed research works.

Table 1. In **perception**, *adversarial attacks* introduce subtle perturbations in sensory inputs, such as visual or auditory data, causing misclassifications and leading to incorrect environmental interpretations. **Backdoor attacks** embed hidden triggers in the model that activate malicious behavior when prompted, while **sensor attacks**, such as spoofing and jamming, compromise sensor data, resulting in environmental sensing failures or system shutdowns. These vulnerabilities can lead to misinterpretation of objects, failure to detect obstacles, or navigation errors. In **cognition**, *adversarial attacks* manipulate reasoning processes, causing the system to make unsafe or incorrect decisions, such as faulty spatial understanding or misinterpretation of context. In **planning**, various attacks, including *adversarial attacks* on task planning and trajectory planning, **jailbreak attacks**, and **backdoor attacks**, can manipulate the model's planned actions, leading to unsafe trajectories, collisions, or failure to follow intended goals. In **action & interaction**, *adversarial manipulations* and **backdoor attacks** can bypass safety mechanisms during human-agent interactions, inducing harmful or unintended behavior, such as violating safety protocols or performing actions that harm users. Finally, in **agentic systems**, threats arise from the agent's expanded autonomy: *tool misuse* can lead to harmful code execution or unintended physical actions, **memory poisoning** can corrupt the agent's experience store to cause persistent unsafe behavior, **memory leakage** can expose private user data and privileged context from agent memory stores, and **cascading failures** can propagate through inner layers when self-evolving agents erode their own alignment. In Table 1, we also categorize the potential real-world dangers caused by these threats.

Differences from Existing Surveys. Prior surveys examine embodied AI safety from complementary perspectives. [397] provides an early analysis of vulnerabilities and attack surfaces but offers limited coverage of defensive strategies. [381] focuses on robustness issues in navigation but does not extend to cognition, manipulation, or human–robot interaction. [324] presents conceptual foundations and system-level safety principles but does not develop detailed attack–defense taxonomies. [24] analyzes world-model safety, particularly predictive failures, while leaving perception, planning, and interaction risks less explored. [247] argues that embodied failures arise from system-level mismatches rather than isolated LLM or CPS flaws, but does not develop component-level attack–defense taxonomies. [366] examines adversarial robustness from a

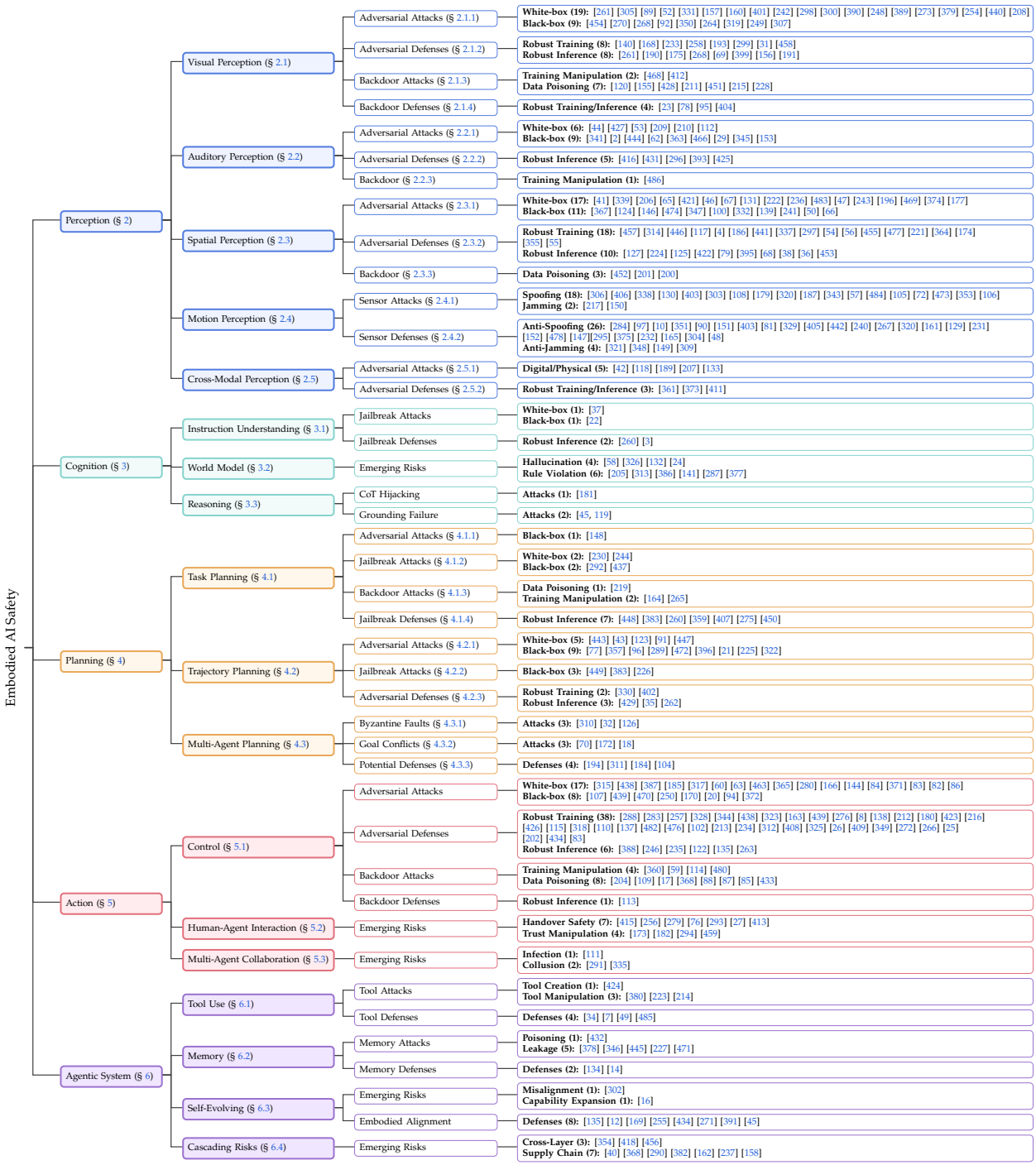

Figure 4 The roadmap of this survey.

closed-loop propagation perspective, but focuses on adversarial attacks without covering backdoor, jailbreak, or agentic-level threats. [143] surveys security threats and defenses for LLM-controlled robotics, but scopes narrowly to LLM integration without addressing broader perception or interaction layers. [281] provides a policy-oriented risk taxonomy spanning physical, informational, and social dimensions, but does not analyze specific attack mechanisms or defenses. [376] surveys embodied AI from an IoT perspective with a dual-brain architecture framework, but treats security and privacy as one component among enabling technologies

Table 1 Summary of attacks and threats across capability layers of embodied AI.

Capability Layer	Attack	Threat	Real-world Danger
Perception	Adversarial Attack	Misclassification, misdetection, scene misinterpretation	Wrong object recognition, traffic sign errors, surveillance failure
	Backdoor Attack	Triggered misperception, hidden model manipulation	Unsafe behavior activation, safety bypass during deployment
	Sensor Attack	Sensor spoofing, jamming, data corruption	Navigation failure, loss of situational awareness, system malfunction
Cognition	Adversarial Attack	Faulty reasoning, scene misunderstanding, context errors	Navigation errors, hazard avoidance failure, unsafe decisions
Planning	Adversarial Attack	Planning perturbation, trajectory errors	Collision risk, unstable motion, unsafe task execution
	Jailbreak Attack	Safety constraint bypass, unsafe goal generation	Execution of prohibited actions, violation of safety rules
	Backdoor Attack	Triggered policy manipulation, hidden planning bias	Malicious plans, safety mechanism bypass
Action & Interaction	Adversarial Attack	Action manipulation, safety guard evasion	Unsafe human–robot interaction, physical injury risk
	Backdoor Attack	Triggered harmful actions, hidden interaction flaws	Malicious responses, loss of control, safety violation
Agentic System	Tool Misuse	Unsafe tool calls, harmful code execution	Physical damage, unsafe API or actuator commands
	Memory Poisoning	Corrupted memory, unsafe policy update	Repeated unsafe behavior, long-term reliability loss
	Memory Leakage	Sensitive memory exposure, data extraction	Privacy breach, leakage of logs or user data
	Cascading Failure	Cross-layer error propagation, alignment drift	System-wide failure, uncontrolled self-evolution

rather than developing attack–defense taxonomies. In contrast, our survey synthesizes attacks and defenses across the entire embodied pipeline and integrates insights from traditional AI safety to provide a unified, mechanism-oriented understanding of embodied AI safety. Figure 4 provides a roadmap of this survey.

In summary, the main contributions of this work are:

- We present a systematic survey of **safety research in embodied AI**, organizing prior work into a coherent multi-level taxonomy covering perception, cognition, planning, action & interaction, and agentic system.
- We review over **400** papers, consolidating embodied-specific research with safety-relevant advances in vision, language, and multimodal foundation models.
- We identify fundamental challenges, open problems, and future research directions, offering a roadmap for developing embodied agents that are not only capable and autonomous but also safe and trustworthy in real-world environments.

2 Perception

Perception forms the innermost layer of embodied AI, granting agents the foundational capability to interpret their environment through multiple sensing modalities. At this layer, the attack surface originates at the sensory boundary: adversaries can corrupt what the agent perceives through adversarial perturbations, sensor spoofing, and backdoor triggers. Because perception underpins all outer layers, errors at this stage

propagate and amplify throughout the system: a misclassified object leads to flawed reasoning, unsafe plans, and dangerous actions. Each modality interfaces with the physical world through distinct sensors and signal types, producing largely modality-specific threat models, attack techniques, and defenses. This section organizes perception by sensing modality: **Visual Perception** (Section 2.1) addresses vulnerabilities in camera-based perception, with emphasis on modern visual encoders (e.g., CLIP, ViT, and SigLIP) used in vision-language models; **Auditory Perception** (Section 2.2) covers attacks on speech recognition and speaker verification systems critical for voice-based human-agent interaction; **Spatial Perception** (Section 2.3) examines threats to 3D understanding including SLAM, depth estimation, pose estimation, and neural scene representations (e.g., NeRF and 3DGS); **Motion Perception** (Section 2.4) addresses vulnerabilities in inertial measurement, GPS, and proprioceptive sensing; and **Cross-Modal Perception** (Section 2.5) discusses attacks on multimodal perception systems and sensor fusion. For each modality, we review attacks, defenses, and evaluation benchmarks. Sensor-level attacks (spoofing and jamming) are integrated within each modality rather than treated separately.

2.1 Visual Perception

Visual perception encompasses camera-based tasks such as object classification, object detection (including lane detection), object tracking, semantic segmentation, and video understanding (action recognition, optical flow, and video object segmentation), each critical for embodied downstream tasks. Modern visual encoders, including contrastive vision-language models (e.g., CLIP and SigLIP) and Vision Transformers (ViT), serve as shared perception backbones whose vulnerabilities propagate to all downstream systems. This subsection consolidates all visual perception security research: adversarial attacks and defenses that manipulate or protect pixel-level inputs, as well as backdoor attacks and defenses that embed or remove hidden triggers in visual models. We organize the discussion into four parts: **Adversarial Attacks** (Section 2.1.1), **Adversarial Defenses** (Section 2.1.2), **Backdoor Attacks** (Section 2.1.3), and **Backdoor Defenses** (Section 2.1.4).

2.1.1 Adversarial Attacks

Adversarial attacks on vision pipelines typically occur in two domains: in the digital space, where they perturb pixel values, and in the physical world, where they manipulate real-world signals (e.g., road signs and flashlights) to deceive perception systems.

White-box Attacks. White-box attacks exploit full model access to craft precise perturbations, organized into digital and physical attack strategies. Digital attacks manipulate inputs directly in the digital space. For object classification, Melis et al. [261] introduced region-constrained perturbations against the iCub robot's vision pipeline. For object detection, Thys et al. [331] used adversarial patches to conceal detections or degrade localization. Adversarial Overlay [390] proposes real-time attacks, and HitM [389] introduces a human-in-the-middle threat model that intercepts camera data before OS processing. For single-object tracking, RTAA [157] exploits temporal information by leveraging motion and recent predictions across frames. TrackPGD [273] targets Transformer-based trackers specifically. For multi-object tracking, under tracking-by-detection (TBD) and joint-detection-tracking (JDT) paradigms, Tracker Hijacking [160] exploits sparse-frame perturbations to induce long-term tracking failures. Ma et al.'s attack [248] corrupts the detection stage via adversarial patches. For semantic segmentation, CAA [379] performs multi-task attacks on joint networks. Uncertainty [254] applies loss-weighting based on uncertainty.

Modern visual encoders introduce new attack surfaces beyond task-specific models. For contrastive vision-language encoders such as CLIP and SigLIP, whose compromise cascades to all downstream VLMs and embodied agents, AnyAttack [440] pre-trains a self-supervised perturbation generator on LAION-400M that produces cross-model attack vectors effective against CLIP, BLIP, BLIP2, and commercial systems without label supervision.

Video perception models face temporal adversarial threats absent from single-image models. PCFA [300] targets optical flow models with global perturbations that shift predicted flow toward attacker-chosen targets.

Table 2 A summary of **adversarial attacks** on **visual perception**.

Attack	Method	Year	Category	Subcategory	Target Model	Dataset
Adversarial Attack	Melis et al.[261]	2017	White-box	Digital Attack	Object Classifier	iCubWorld
	Thys et al.[331]	2019	White-box	Digital Attack	Object Detector	Inria
	Adversarial Overlay[390]	2023	White-box	Digital Attack	Object Detector	PASCAL VOC, ROS Gazebo
	HitM[389]	2024	White-box	Digital Attack	Object Detector	CARLA, VOC
	RTAA[157]	2020	White-box	Digital Attack	Single-Object Tracker	OTB, UAV, VOT
	TrackPGD[273]	2024	White-box	Digital Attack	Single-Object Tracker	DAVIS, GOT-10k, UAV, VOT
	Tracker Hijacking[160]	2020	White-box	Digital Attack	Multi-Object Tracker	BDD
	Ma et al.[248]	2023	White-box	Digital Attack	Multi-Object Tracker	BDD
	CAA[379]	2024	White-box	Digital Attack	Semantic Segmentation	Cityscapes, RainCityscapes
	Uncertainty[254]	2024	White-box	Digital Attack	Semantic Segmentation	Cityscapes, VOC
	AnyAttack[440]	2025	White-box	Digital Attack	CLIP Image Encoder	LAION-400M
	PCFA[300]	2022	White-box	Digital Attack	Optical Flow Estimator	Sintel, KITTI
	DARTS[305]	2018	White-box	Physical Attack	Object Classifier	GTSDB, GTSRB
	RP2[89]	2018	White-box	Physical Attack	Object Classifier	GTSRB, LISA
	ShapeShifter[52]	2018	White-box	Physical Attack	Object Detector	Real-world data
	AdvT[401]	2020	White-box	Physical Attack	Object Detector	Real-world data
	SLAP[242]	2021	White-box	Physical Attack	Object Detector	Real-world data
	DRP[298]	2021	White-box	Physical Attack	Object Detector	Comma2k19, LGSVL
	MFDA[208]	2025	White-box	Physical Attack	Single-Object Tracker	CARLA
	AdvTraj[350]	2024	Black-box	Digital Attack	Multi-Object Tracker	CARLA
	Fang et al.[92]	2023	Black-box	Digital Attack	Object Detector	Carlasc, Comma2k19, CULane
	PB-UAP[307]	2025	Black-box	Digital Attack	Semantic Segmentation	VOC, Cityscapes
	ELA[319]	2024	Black-box	Physical Attack	Object Classifier	CARLA
	CAMOU[454]	2018	Black-box	Physical Attack	Object Detector	Unreal Engine
	MobilBye[270]	2019	Black-box	Physical Attack	Object Detector	Real-world data
	SSPA[268]	2020	Black-box	Physical Attack	Object Detector	Web data
	NS Attack[264]	2024	Black-box	Physical Attack	Object Detector	CARLA
	ControlLoc[249]	2024	Black-box	Physical Attack	Multi-Object Tracker	BDD, KITTI

Table 3 A summary of **adversarial defenses** for **visual perception**.

Defense	Method	Year	Category	Subcategory	Target Model	Dataset
Adversarial Defense	Kalin et al.[168]	2021	Robust Training	Adversarial Training	Object Classifier	VEDAI
	DSNet[140]	2020	Robust Training	Adversarial Training	Object Detector	FOD, Foggy Driving
	IA-YOLO[233]	2022	Robust Training	Adversarial Training	Object Detector	VOC_Foggy, RTTS
	BAD-Net[193]	2023	Robust Training	Adversarial Training	Object Detector	RTTS, VOChaze
	Blazevic et al.[31]	2025	Robust Training	Adversarial Training	Object Detector	MetaDrive
	RP-PGD[458]	2025	Robust Training	Adversarial Training	Semantic Segmentation	Seg-ADE20K, VOC, Cityscapes
	Melis et al.[261]	2017	Robust Inference	Input Moderation	Object Classifier	iCubWorld
	AOD-Net[190]	2018	Robust Inference	Input Moderation	Object Detector	Middlebury, Real-world data
	DGFN[175]	2018	Robust Inference	Input Moderation	Object Detector	KITTI
	GhostBusters[268]	2020	Robust Inference	Input Moderation	Object Detector	Real-world data
	Jia et al.[156]	2024	Robust Inference	Input Moderation	Single-Object Tracker	UAV
	TeCoA[258]	2023	Robust Training	Adversarial Training	CLIP Image Encoder	ImageNet, 15 ZS datasets
	Robust CLIP[299]	2024	Robust Training	Adversarial Training	CLIP Image Encoder	ImageNet, COCO
	SentiNet[69]	2020	Robust Inference	Output Moderation	Object Detector	ImageNet, LFW, LISA, VGG-Face
	Xu et al.[399]	2021	Robust Inference	Output Moderation	Object Detector	TuSimple
	Li et al.[191]	2025	Robust Inference	Output Moderation	Single-Object Tracker	LaSOT, OTB, UAV

Physical attacks modify objects or scenes to deceive perception under realistic capture conditions. For object classification, DARTS [305] and RP2 [89] perform physical attacks by attaching stickers or printing patterns on objects. For object detection, ShapeShifter [52] extended Expectation over Transformation (EoT) to the physical domain. AdvT [401], SLAP [242] and DRP [298] perform physical adversarial attacks on objects or environments by printing patterns on garments, projecting textures onto traffic signs, or disguising patches as road stains. For single-object tracking, MFDA [208] fools SOT models under viewpoint, deformation, and illumination changes.

Black-box Attacks. Black-box attacks operate without model access, relying on transferability, query-based optimization, or surrogate models, organized into digital and physical attack strategies. In the digital domain, for multi-object tracking, AdvTraj [350] confuses the association phase by swapping the attacker's ID with a target's ID. For object detection, Fang et al. [92] used Particle Swarm Optimization to perform heuristic searches on lane-like perturbations for lane detection. For semantic segmentation, PB-UAP [307] generates black-box universal perturbations transferable across models.

In the physical domain, for object classification, ELA [319] predicts traffic-sign poses and trains an RL agent to project adversarial laser patterns in real time. For object detection, CAMOU [454] perturbs vehicle appearances, while MobilBye [270] and SSPA [268] employ optical-based attacks via projecting phantom objects. NS Attack [264] perturbs road appearances to evade detection. For multi-object tracking, ControlLoc [249] searches for optimal patch placements to manipulate objects' positions and shapes.

2.1.2 Adversarial Defenses

Visual defenses protect object classification, object detection (including lane detection), tracking, and segmentation pipelines from adversarial manipulation through robust training and inference strategies.

Robust Training. Robust training hardens models by incorporating adversarial examples, augmented data, or feature recovery during training. For object classification, Kalin et al. [168] retrained models with

visible-light and infrared imagery, guided by adversarial-surface analysis. For object detection, DSNet [140], IA-YOLO [233], and BAD-Net [193] jointly learn visibility enhancement, feature restoration, or dehazing with detection, and Blazevic et al. [31] trained robust lane detection models against adversarial perturbations. For semantic segmentation, RP-PGD [458] employs adversarial training to enhance model robustness.

Robust Inference. Robust inference defends models through input or output moderation. Input moderation detects anomalous inputs, preprocesses signals, restores degraded data, or fuses multimodal information. For object classification, Melis et al. [261] detected and filtered inputs deviating from training distributions in deep feature space. For object detection, AOD-Net [190] provides lightweight dehazing to restore visibility. DGFN [175] fuses camera and LiDAR data with gating mechanisms. GhostBusters [268] deploys four specialized CNNs analyzing various visual features. For single-object tracking, Jia et al. [156] detected attacks using similarity differences in the feature space. Modern visual encoders also require architecture-aware defenses. For CLIP-family encoders, TeCoA [258] introduces contrastive adversarial fine-tuning that improves zero-shot robustness. Robust CLIP [299] proposes unsupervised adversarial fine-tuning to ensure that downstream tasks inherit encoder-level robustness. Output moderation verifies model predictions by analyzing output behavior or comparing predictions across transformations. For object detection, SentiNet [69] employs output-based detection by localizing suspicious regions, while Xu et al. [399] applied a CNN to the detected lanes to classify them as real or fake. For single-object tracking, Li et al. [191] compared full- and low-frequency tracking, using the low-frequency branch as a stable reference.

2.1.3 Backdoor Attacks

Backdoor attacks on visual perception embed hidden triggers during model training so that the system behaves normally on benign inputs but executes attacker-chosen behaviors when the trigger appears. These attacks target visual models through training manipulation or data poisoning.

Training Manipulation. Training manipulation attacks embed backdoors by manipulating training objectives or procedures. For ViTs, TrojViT [468] injects trojans via RowHammer-based bit-flipping without training-time poisoning, and SWARM [412] targets prompt-tuned ViTs with a switchable backdoor.

Data Poisoning. Data poisoning attacks embed backdoors by injecting triggered samples into training data. Han et al. [120] inserted physical objects as triggers using poison-annotation strategies for object detectors. BadLANE [451] and DBALD [215] embed visual pattern triggers via meta-learning or diffusion-based synthesis for object detectors.

Modern visual encoders are also vulnerable to backdoor attacks. For vision-language encoders, a single compromised encoder propagates backdoor behavior to all downstream tasks: BadEncoder [155] first demonstrates this supply-chain threat on pre-trained encoders including CLIP, BadCLIP [211] optimizes triggers via dual-embedding guided Bayesian reasoning, and BadVision [228] exploits SSL encoder backdoors to induce visual hallucinations in LVLMs. For ViTs, BadViT [428] shows that self-attention makes ViTs more sensitive to patch-wise triggers than CNNs.

2.1.4 Backdoor Defenses

Backdoor defenses for visual perception aim to detect or remove hidden visual triggers implanted during training. Current defense research for task-specific visual backdoor attacks remains limited, highlighting an important open challenge for securing visual perception in embodied systems.

For modern visual encoders, backdoor defenses must operate at the encoder level. For ViTs, Doan et al. [78] exploited distinctive patch-transformation responses to detect triggers without training data access. For CLIP-family encoders, CleanCLIP [23] re-aligns modality representations via multimodal contrastive fine-tuning to weaken backdoor associations, DECREE [95] detects backdoors in pre-trained encoders without classifier headers or input labels, and BDetCLIP [404] enables efficient test-time backdoor detection via contrastive prompting.

2.2 Auditory Perception

Auditory perception supports human-robot interaction and voice-based control through speech recognizers, which convert spoken language into text, and speaker verifiers, which authenticate speaker identity based on voice characteristics. This subsection consolidates all auditory perception security research, organized into three parts: **Adversarial Attacks** (Section 2.2.1), **Adversarial Defenses** (Section 2.2.2), and **Backdoor Attacks and Defenses** (Section 2.2.3).

2.2.1 Adversarial Attacks

Adversarial attacks on auditory perception operate in the digital space by perturbing audio waveforms and in the physical space by manipulating signals (e.g., voice injection and speaker spoofing).

White-box Attacks. In the physical space, for speech recognizers, Carlini et al. [44] converted audio commands into forms unintelligible to humans. CommanderSong [427] embeds perturbations into songs. Metamorph [53] employs background-like audio perturbations, while SpecPatch [112] employs audio spectrogram patches. For speaker verifiers, Li et al. [209] incorporated Room Impulse Response to maintain effectiveness under over-the-air playback. For the joint speech recognizers and speaker verifiers, AdvPulse [210] designs subsecond perturbations.

Black-box Attacks. In the digital space, TSMAE [153] adjusts playback speed to attack speech recognizers. Occam [466] introduces decision-only adversarial examples for cloud APIs. In the physical space, for speech recognizers, BarrierBypass [345] injects commands through physical barriers. Cocaine Noodles [341] and Abdullah et al. [2] generated mangled inputs that are unintelligible to humans but still recognized by machines. Devil's Whisper [62] crafts adversarial examples by pairing a query-based substitute with a stronger white-box recognizer. Wang et al. [363] modulated the signals to compensate for the distortion in the frequency domain. For speaker verifiers, Vaspy [444] synthesizes activation keywords via speech recognition and voice cloning. Bilika et al. [29] demonstrated practical feasibility in real-world scenarios.

2.2.2 Adversarial Defenses

Auditory defenses protect speech recognizer and speaker verifier systems from adversarial perturbations through robust inference strategies.

Robust Inference. For input moderation of speech recognizers, Samizade et al. [296] extracted MFCC features and classified benign and adversarial samples using CNNs. Yang et al. [416] explored input preprocessing techniques, including perturbation, compression, quantization, smoothing, reconstruction, and downsampling. AudioPure [393] employs diffusion models to purify and restore input signals. For speaker verifiers, AntiFake [425] embeds protective perturbations into audio to prevent voice cloning and forgery. For output moderation of speech recognizers, Yang et al.'s approach [416] compares transcription consistency between audio segments and complete audio. MVP-EARS [431] uses multiple models for cross-verification.

2.2.3 Backdoor Attacks and Defenses

Backdoor research on auditory perception remains nascent. TrojanModel [486] inserts backdoors into acoustic models for speech recognizers via training manipulation, but no dedicated defense has been proposed, highlighting an important open challenge for securing speech-based embodied systems.

2.3 Spatial Perception

Spatial perception enables embodied agents to build, maintain, and reason about 3D representations of their environment for navigation, manipulation, and obstacle avoidance. It encompasses point cloud classification, 3D object detection and tracking, trajectory prediction, depth and pose estimation, SLAM, and neural scene representations (NeRF, 3DGS). This subsection consolidates all spatial perception security research, organized

Table 4 A summary of **adversarial** attacks and defenses for **auditory perception**.

Attack/Defense	Method	Year	Category	Subcategory	Target Model	Dataset
Adversarial Attack	Carlini et al.[44]	2016	White-box	Physical Attack	Speech Recognizer	Custom dataset
	CommanderSong[427]	2018	White-box	Physical Attack	Speech Recognizer	Custom dataset
	Metamorph[53]	2020	White-box	Physical Attack	Speech Recognizer	AIR, Common Voice, MARDY
	SpecPatch[112]	2022	White-box	Physical Attack	Speech Recognizer	TIMIT
	Li et al.[209]	2020	White-box	Physical Attack	Speech Verifier	CSTR VCTK Corpus
	AdvPulse[210]	2020	White-box	Physical Attack	Speech Rec./Ver.	CSTR VCTK Corpus, Voice Commands
	TSMAE[153]	2024	Black-box	Digital Attack	Speech Recognizer	CSTR VCTK Corpus, Custom dataset
	Occam[466]	2021	Black-box	Digital Attack	Speech Rec./Ver.	Common Voice, LibriSpeech, VoxCeleb
	Cocaine Noodles[341]	2015	Black-box	Physical Attack	Speech Recognizer	Custom dataset
	Wang et al.[363]	2020	Black-box	Physical Attack	Speech Recognizer	Custom dataset
	Devil's Whisper[62]	2020	Black-box	Physical Attack	Speech Recognizer	CommanderSong dataset
	BarrierBypass[345]	2023	Black-box	Physical Attack	Speech Recognizer	Custom dataset
	Vaspy[444]	2019	Black-box	Physical Attack	Speech Verifier	Custom dataset
	Bilika et al.[29]	2023	Black-box	Physical Attack	Speech Verifier	Custom dataset
	Abdullah et al.[2]	2019	Black-box	Physical Attack	Speech Rec./Ver.	LibriSpeech, TIMIT
Adversarial Defense	Samizade et al.[296]	2020	Robust Inference	Input Moderation	Speech Recognizer	Common Voice, Speech Commands
	Yang et al.[416]	2018	Robust Inference	Input Moderation	Speech Recognizer	Common Voice, LIBRIS
	AudioPure[393]	2023	Robust Inference	Input Moderation	Speech Recognizer	Qualcomm Keyword Speech
	AntiFake[425]	2023	Robust Inference	Input Moderation	Speech Verifier	CSTR VCTK, LibriSpeech, TIMIT
	Yang et al.[416]	2018	Robust Inference	Output Moderation	Speech Recognizer	Common Voice, LIBRIS
	MVP-EARS[431]	2019	Robust Inference	Output Moderation	Speech Recognizer	Common Voice, Custom dataset

into three parts: **Adversarial Attacks** (Section 2.3.1), **Adversarial Defenses** (Section 2.3.2), and **Backdoor Attacks and Defenses** (Section 2.3.3).

2.3.1 Adversarial Attacks

Adversarial attacks on spatial perception models perturb point clouds, depth maps, or neural scene representations in the digital space, or manipulate 3D object geometries, LiDAR signals, or camera inputs in the physical space to disrupt detection, localization, and navigation.

White-box Attacks. White-box attacks exploit full model access to craft precise perturbations, organized into digital and physical attack strategies. In the digital domain, for 3D object detectors, FLAT [206] manipulates the vehicle trajectory to affect LiDAR motion compensation. SlowLiDAR [222] introduces slow-acting perturbations, causing delayed failures. Zheng et al. [469] leveraged smoke-like perturbations to interfere with sensing, and Wang et al. [374] utilized saliency maps to identify critical points before optimizing perturbations. For 3D object trackers, Cheng et al.'s method [65] crafts universal perturbations, and TAN [236] crafts transferable perturbations. For SLAM systems, Yoshida et al. [421] applied small, well-timed perturbations to LiDAR point clouds to disrupt scan matching and degrade map consistency, leading to accumulating localization errors. For scene representation models, Horváth and Józsa [131] demonstrated that NeRFs can be subverted by carefully perturbed input views to render photorealistic but falsified scenes, and Poison-Splat [243] introduces a data-poisoning attack on 3DGS that corrupts adaptive density control, causing denial-of-service through excessive memory and compute usage. In the physical domain, for 3D object detectors, LiDAR-Adv [41] and Tu et al.'s attack [339] optimize adversarial 3D mesh geometries under physical constraints. AE-Morpher [483] generates adversarial meshes with morphing constraints. Adv3D [196] embeds adversarial objects directly as NeRFs, enabling transferable and contact-free perturbations. ShadowHack [177] exploits optimized planar materials to manipulate shadow patterns. For depth estimation, Cheng et al. [67] optimized physical adversarial patches that bias monocular depth estimators. For pose estimation, Chawla et al. [46] crafted patches that yield large trajectory errors. For visual SLAM, AoR [47] uses adversarial patches to trigger false loop closures, producing severe localization drift.

Black-box Attacks. In the digital space, for 3D object detectors, Hau et al. [124] forced the sensor to record injected points, thereby pushing the genuine points outside the object's bounding box. For 3D object trackers, TAPG [332] and Cheng et al. [66] optimized black-box attacks via transfer-based and explainability-guided methods, respectively. For NeRF-based navigation, Wang et al. [347] benchmarked NeRF navigators under visual corruptions. In the physical domain, for 3D object detectors, SpotAttack [139] employs a genetic algorithm-based global search to optimize non-reflective adversarial spots, while LiDAttack [50] combines global search with local refinement for covert attacks. For camera-based systems, ICSL Attack [367] creates ghost traffic signals using infrared projection invisible to humans. For stereo depth estimation, DoubleStar [474] uses long-range, synchronized light patterns to exploit stereo-matching artifacts and fabricate obstacles for drones. For visual SLAM, Ikram et al. [146] employed duplicated textured regions to trigger perceptual aliasing and induce localization drift. For LiDAR SLAM, Fukunaga et al. [100] injected simple, well-timed points into LiDAR streams to disrupt scan matching. For trajectory predictors, Lou et al. [241] introduced the first physical-world attack on trajectory prediction via LiDAR-induced deceptions, employing a two-stage framework that identifies velocity-insensitive state perturbations and matches them to feasible object locations.

2.3.2 Adversarial Defenses

Spatial defenses protect point cloud classifiers, 3D object detectors, SLAM systems, and neural scene representations from adversarial perturbations through robust training and robust inference strategies.

Robust Training. Robust training strengthens spatial perception models by exposing them to adversarial examples, augmented data, or physical constraints during the training process. We refer to adversarial training as a general framework where training data is augmented by either natural perturbations (e.g., weather, fog, crash simulations) or adversarial perturbations (e.g., gradient-based attacks) to improve model robustness

Table 5 A summary of **adversarial attacks** on **spatial perception**.

Attack	Method	Year	Category	Subcategory	Target Model	Dataset
Adversarial Attack	FLAT[206]	2021	White-box	Digital Attack	3D Object Detector	nuScenes
	SlowLiDAR[222]	2023	White-box	Digital Attack	3D Object Detector	KITTI
	Zheng et al.[469]	2025	White-box	Digital Attack	3D Object Detector	KITTI, nuScenes
	Wang et al.[374]	2025	White-box	Digital Attack	3D Object Detector	KITTI, Waymo
	Cheng et al.[65]	2021	White-box	Digital Attack	3D Object Tracker	KITTI
	TAN[236]	2023	White-box	Digital Attack	3D Object Tracker	KITTI
	Yoshida et al. [421]	2022	White-box	Digital Attack	SLAM (LiDAR)	Self-constructed data
	Horváth and Józsa [131]	2023	White-box	Digital Attack	NeRF Navigator	LLFF
	Poison-Splat [243]	2024	White-box	Digital Attack	3DGS Navigator	NeRF-Synthetic, Mip-NeRF360
	LiDAR-Adv[41]	2019	White-box	Physical Attack	3D Object Detector	Real-world data
	Tu et al.[339]	2020	White-box	Physical Attack	3D Object Detector	KITTI
	AE-Morpher[483]	2024	White-box	Physical Attack	3D Object Detector	Custom dataset, SVL simulator
	Adv3D [196]	2024	White-box	Physical Attack	3D Object Detector	nuScenes
	ShadowHack[177]	2025	White-box	Physical Attack	3D Object Detector	AWSIM simulator
	Cheng et al. [67]	2022	White-box	Physical Attack	Depth Estimator	KITTI
	Chawla et al. [46]	2022	White-box	Physical Attack	Pose Estimator	KITTI odometry
	AoR [47]	2024	White-box	Physical Attack	SLAM (Visual)	KITTI, Oxford RobotCar, 4Seasons
	Hau et al.[124]	2021	Black-box	Digital Attack	3D Object Detector	KITTI
	TAPG[332]	2024	Black-box	Digital Attack	3D Object Tracker	KITTI Tracking, nuScenes
	Cheng et al.[66]	2025	Black-box	Digital Attack	3D Object Tracker	KITTI, nuScenes, Waymo
	Wang et al. [347]	2024	Black-box	Digital Attack	NeRF Navigator	LLFF-C, Blender-C
	SpotAttack[139]	2024	Black-box	Physical Attack	3D Object Detector	MATLAB simulator
	LiDAttack[50]	2025	Black-box	Physical Attack	3D Object Detector	Custom dataset, KITTI, nuScenes
	ICSL Attack [367]	2021	Black-box	Physical Attack	Camera	Bosch Night, KITTI
	DoubleStar [474]	2022	Black-box	Physical Attack	Stereo Depth	Self-constructed data
	Ikram et al. [146]	2022	Black-box	Physical Attack	SLAM (Visual)	Manhattan, Intel, MIT, Garage
	Fukunaga et al. [100]	2024	Black-box	Physical Attack	LiDAR SLAM	Self-constructed data
	Lou et al. [241]	2024	Black-box	Physical Attack	Trajectory Predictor	nuScenes, Real world

Table 6 A summary of **adversarial defenses** for **spatial perception**.

Defense	Method	Year	Category	Subcategory	Target Model	Dataset
Adversarial Defense	Defense-PointNet[457]	2019	Robust Training	Adversarial Training	Point Cloud Classifier	ShapeNet
	Sun et al.[314]	2021	Robust Training	Adversarial Training	Point Cloud Classifier	ModelNet, ScanObjectNN
	PointCutMix[441]	2022	Robust Training	Adversarial Training	Point Cloud Classifier	ModelNet, ScanObjectNN
	Hahner et al.[117]	2021	Robust Training	Adversarial Training	3D Object Detector	KITTI
	3D-VField[186]	2022	Robust Training	Adversarial Training	3D Object Detector	CrashD, KITTI, Waymo
	BAFT[455]	2024	Robust Training	Adversarial Training	3D Object Detector	KITTI, Waymo
	DART[355]	2025	Robust Training	Adversarial Training	3D Object Detector	KITTI
	LISA[174]	2025	Robust Training	Adversarial Training	3D Object Detector	KITTI, Waymo
	AcousticFusion [446]	2021	Robust Training	Multi-Modal Fusion	SLAM	Azure Kinect audio and RGB-D
	Wang et al. [364]	2024	Robust Training	Multi-Modal Fusion	Safety System	Self-constructed data
	Sang et al. [297]	2023	Robust Training	Scene Augmentation	Scene Understanding	iGibson
	Adamkiewicz et al. [4]	2022	Robust Training	Environment Augmentation	NeRF Navigator	Self-constructed data
	Splat-Nav [55]	2025	Robust Training	Environment Augmentation	3DGS Navigator	Stonehenge, Statues, Flightroom
	Tong et al. [337]	2023	Robust Training	Formal Methods	Safety	NeRF Navigator	Replica Dataset
	CATNIPS [54]	2024	Robust Training	Formal Methods	Safety	Navigation	Stonehenge, Statues, Flightroom
	Zhou et al. [477]	2024	Robust Training	Formal Methods	Safety	SLAM + Control	Self-constructed data
	SAFER-Splat [56]	2024	Robust Training	Formal Methods	Safety	3DGS Navigator	Self-constructed data
	RaEM [221]	2024	Robust Training	Risk-Aware Planning	Active Perception	Matterport3D
	PointGuard[224]	2021	Robust Inference	Input Moderation	Point Cloud Classifier	ModelNet40, ScanNet
	WeatherNet[127]	2020	Robust Inference	Input Moderation	3D Object Detector	Chamber & road(custom)
	Shadow-Catcher[125]	2021	Robust Inference	Input Moderation	3D Object Detector	KITTI
	LOP[395]	2023	Robust Inference	Input Moderation	3D Object Detector	KITTI, LGSVL simulator
	ADoPT[68]	2023	Robust Inference	Input Moderation	3D Object Detector	nuScenes
	LiDARPure[38]	2024	Robust Inference	Input Moderation	3D Object Detector	KITTI
	Zhang et al. [453]	2025	Robust Inference	Input Moderation	3D Object Detector	nuScenes, KITTI
	Brunke et al. [36]	2025	Robust Inference	Input Moderation	Scene Understanding	ScanNet200, Self-constructed data
	3D-TC2[422]	2021	Robust Inference	Output Moderation	3D Object Detector	nuScenes
	ViewFool [79]	2022	Robust Inference	Output Moderation	Object Classifier	BlenderKit, Objectron

under distribution shift and adversarial manipulation. For point cloud classifiers, Defense-PointNet [457], Sun et al.'s work [314], and PointCutMix [441] employ robust training with various data augmentation or perturbation strategies. For 3D object detectors, Hahner et al. [117] introduced LiDAR fog simulation and augmentation techniques. 3D-VField [186] learns vector fields for adversarial augmentation, BAFT [455] and DART [355] use optimized perturbation strategies, and LISA [174] further improves robustness via LiDAR-specific augmentation.

Beyond adversarial training, spatial perception benefits from multi-modal fusion, scene augmentation, environment augmentation, formal safety methods, and risk-aware planning. AcousticFusion [446] fuses auditory cues with visual SLAM to stabilize localization under motion and scene changes, and Wang et al. [364] developed a cooperative safety system fusing multi-camera 3D inputs for dynamic human detection and real-time safety-zone enforcement. Sang et al. [297] proposed systematic scene augmentation that procedurally varies layouts, objects, and states to improve generalization. Adamkiewicz et al. [4] achieved collision-free navigation in NeRF-modeled environments, and Splat-Nav [55] demonstrate real-time safe navigation using Gaussian Splatting. Tong et al. [337] paired predictive NeRF rendering with CBF filtering, CATNIPS [54] reinterprets NeRF densities for rigorous collision-probability estimation, Zhou et al. [477] coupled visual-inertial SLAM with control barrier functions (CBFs), and SAFER-Splat [56] embeds CBFs directly over Gaussian Splatting primitives. RaEM [221] introduces risk-aware view acquisition that prioritizes safety-critical regions.

Robust Inference. Input moderation for spatial perception models detects anomalous inputs, preprocesses point clouds, or restores degraded data before inference. For point cloud classifiers, PointGuard [224] provides certified robustness guarantees. For 3D object detectors, WeatherNet [127] is trained to remove noise induced by bad weather. Shadow-Catcher [125] and LOP [395] detect adversarial point clouds via physical invariants such as 3D shadows and depth-density relations. ADoPT [68] leverages temporal consistency for abnormal input detection. LiDARPure [38] employs diffusion models to purify input point clouds. Zhang et al. [453] proposed the first real-time defense against object-based LiDAR attacks, employing a generative model positioned between sensing and perception to identify and remove adversarial points from suspicious regions in point clouds. For scene understanding, Brunke et al. [36] introduced a semantic safety filter integrating 3D semantic maps with LLM reasoning, compiling abstract language-grounded rules into CBFs that enforce both geometric and semantic safety. For output moderation, 3D-TC2 [422] verifies predictions of 3D object detectors via temporal consistency across frames. ViewFool [79] employs NeRF to systematically identify failure-inducing viewpoints for robustness assessment of object classifiers.

2.3.3 Backdoor Attacks and Defenses

Backdoor attacks on spatial perception target 3D object detectors and LiDAR segmentation through data poisoning. Zhang et al. [452] and BadLiDet [201] demonstrate pixel-level trigger injection in bird's-eye-view representations and imperceptible point-level perturbations for 3D object detectors, while BadLiSeg [200] embeds backdoors in LiDAR segmentation via crafted spoofed patterns. No dedicated defense has been proposed for spatial backdoors, an important open challenge for securing 3D perception in embodied systems.

2.4 Motion Perception

Motion perception enables embodied agents to estimate their own pose, velocity, and trajectory through inertial measurement units (IMUs), visual and LiDAR odometry, Global Navigation Satellite System (GNSS) receivers, and simultaneous localization and mapping (SLAM) pipelines. Compromised motion perception leads directly to dangerous physical behavior: erroneous position estimates cause collision, drift, or loss of navigation, while corrupted pose estimation destabilizes control loops in drones, ground robots, and autonomous vehicles. We organize the discussion into two parts: **Sensor Attacks** (Section 2.4.1) cover IMU-based perception attacks, localization and odometry attacks, and sensor-level spoofing and jamming of GNSS, IMU, ultrasonic, and mmWave radar systems; and **Sensor Defenses** (Section 2.4.2) protect motion estimation through anomaly detection, cross-sensor verification, robust state estimation, and anti-spoofing/anti-jamming

Table 7 A summary of **backdoor** attacks and defenses for **embodied perception**.

Attack/Defe	Method	Year	Category	Subcategory	Target Model	Dataset
Backdoor Attack	TrojViT[468]	2023	Training Manipulation	Bit-Flipping	Vision Transformer	ImageNet
	SWARM[412]	2024	Training Manipulation	Prompt Poisoning	Vision Transformer	CIFAR-100, ImageNet
	TrojanModel[486]	2023	Training Manipulation	Trigger Injection	Speech Recognizer	Google Speech Commands
	Han et al.[120]	2022	Data Poisoning	Physical Object	Object Detector	TuSimple
	BadLANE[451]	2024	Data Poisoning	Visual Pattern	Object Detector	CULane, TuSimple
	DBALD[215]	2025	Data Poisoning	Visual Pattern	Object Detector	CULane, TuSimple
	BadEncoder[155]	2022	Data Poisoning	Embedding Poisoning	CLIP Image Encoder	CIFAR-10, STL-10, SVHN
	BadCLIP[211]	2024	Data Poisoning	Embedding Poisoning	CLIP Image Encoder	ImageNet, 11 ZS datasets
	BadVision[228]	2025	Data Poisoning	Embedding Poisoning	CLIP Image Encoder	COCO, VQA-v2
	BadViT[428]	2023	Data Poisoning	Attention Manipulation	Vision Transformer	CIFAR-10, ImageNet
	Zhang et al.[452]	2022	Data Poisoning	BEV Trigger	3D Object Detector	KITTI
	BadLiDet[201]	2023	Data Poisoning	Point Perturbation	3D Object Detector	KITTI, nuScenes
	BadLiSeg[200]	2023	Data Poisoning	Spoofed Pattern	3D Segmentation	SemanticKITTI
Backdoor Defense	Doan et al.[78]	2023	Robust Inference	Patch Processing	Vision Transformer	CIFAR-10, ImageNet
	CleanCLIP[23]	2023	Robust Training	Fine-Tuning	CLIP Image Encoder	ImageNet, 8 ZS datasets
	DECREE[95]	2023	Robust Inference	Detection	CLIP Image Encoder	CIFAR-10, ImageNet, STL-10
	BDetCLIP[404]	2025	Robust Inference	Test-Time Detection	CLIP Image Encoder	ImageNet, 11 ZS datasets

Table 8 A summary of **sensor attacks** on **motion perception**. RF: Radio Frequency; EM: Electromagnetic.

Attack	Method	Year	Category	Subcategory	Target Sensor	Dataset
Sensor Attack	Lenhart et al.[187]	2021	Spoofing	Replay Spoofing	GNSS	real-world data
	Wang et al.[353]	2025	Spoofing	Replay Spoofing	GNSS	ESA, real-world data
	Horton et al.[130]	2018	Spoofing	Generative Spoofing	GNSS	real-world data
	FusionRipper[303]	2020	Spoofing	Generative Spoofing	GNSS	Apollo Data, KAIST Complex Urban
	Dasgupta et al.[72]	2024	Spoofing	Generative Spoofing	GNSS	custom dataset
	Zhong et al.[473]	2025	Spoofing	Generative Spoofing	GNSS	real-world data
	Son et al.[306]	2015	Spoofing	Acoustic Injection	IMU	real-world data
	WALNUT[338]	2017	Spoofing	Acoustic Injection	IMU	real-world data
	KITE[105]	2023	Spoofing	Acoustic Injection	IMU	custom dataset, real-world data
	Yan et al.[406]	2016	Spoofing	Acoustic Injection	Ultrasonic Ranging	real-world data
	Xu et al.[403]	2018	Spoofing	Acoustic Injection	Ultrasonic Ranging	real-world data
	Gluck et al.[108]	2020	Spoofing	Acoustic Injection	Ultrasonic Ranging	real-world data
	Sun et al.[320]	2021	Spoofing	RF Injection	mmWave Radar	real-world data
	Komissarov et al.[179]	2021	Spoofing	RF Injection	mmWave Radar	real-world data
	mmSpoof[343]	2023	Spoofing	RF Injection	mmWave Radar	real-world data
	MetaWave[57]	2023	Spoofing	Physical Manipulation	mmWave Radar	real-world data
	TileMask[484]	2023	Spoofing	Physical Manipulation	mmWave Radar	real-world data
	mmHide[106]	2025	Spoofing	Physical Manipulation	mmWave Radar	real-world data
	Lim et al.[217]	2018	Jamming	Acoustic Interference	Ultrasonic Ranging	real-world data
	Jang et al.[150]	2023	Jamming	EM Interference	IMU	custom dataset, real-world data

mechanisms.

2.4.1 Sensor Attacks

Sensor attacks on motion perception manipulate raw sensor signals before data reaches perception algorithms, exploiting hardware vulnerabilities, signal-processing pipelines, or physical-layer characteristics. These attacks compromise the integrity of raw measurements through physical injection and interference to achieve spoofing (deception) or jamming (disruption) goals.

Spoofing. Spoofing encompasses adversarial techniques that manipulate sensor inputs by injecting or presenting false, yet physically plausible, signals, thereby inducing erroneous measurements or misleading perception.

For GNSS, replay spoofing captures and retransmits authentic signals with intentional delay. Lenhart et al. [187] demonstrated long-range real-time relay systems using commercial software-defined radios (SDRs), and Wang et al. [353] exposed vulnerabilities in Galileo's OSNMA through artificially manipulated time synchronization.

Generative spoofing synthesizes counterfeit yet realistic GNSS signals. Horton et al. [130] leveraged low-cost SDR platforms to generate spoofing signals; Shen et al. [303] crafted fusion-aware perturbations to mislead integrated navigation systems; Dasgupta et al. [72] devised slow-drift attacks; and Zhong et al. [473] analyzed perturbations against integrated navigation.

Table 9 A summary of **sensor defenses** for **motion perception**. RF: Radio Frequency; EM: Electromagnetic.

Defense	Method	Year	Category	Subcategory	Target Sensor	Dataset
Sensor Defense	Falco et al.[90]	2018	Anti-Spoofing	Detection	GNSS	Simulation data
	Crowd-GPS-Sec[151]	2018	Anti-Spoofing	Detection	GNSS	Real-world data, Simulation data
	DeepSIM[405]	2020	Anti-Spoofing	Detection	GNSS	SatUAV(custom)
	DeepPOSE[161]	2022	Anti-Spoofing	Detection	GNSS	BDD-100K, Custom dataset
	Iqbal et al.[147]	2024	Anti-Spoofing	Detection	GNSS	TEXBAT
	PADS[232]	2025	Anti-Spoofing	Detection	GNSS	Jammertest
	Jin et al.[165]	2025	Anti-Spoofing	Detection	GNSS	Real-world data
	SDI[329]	2020	Anti-Spoofing	Detection	IMU	Custom dataset, Real-world data
	Liu et al.[231]	2022	Anti-Spoofing	Detection	IMU	Custom dataset, Real-world data
	CPD-MhIMU[295]	2024	Anti-Spoofing	Detection	IMU	Custom dataset, Real-world data
	Xu et al.[403]	2018	Anti-Spoofing	Detection	Ultrasonic Ranging	real-world data
	SoundFence[240]	2021	Anti-Spoofing	Detection	Ultrasonic Ranging	real-world data
	SecureTrack[304]	2025	Anti-Spoofing	Detection	Ultrasonic Ranging	real-world data
	Sun et al.[320]	2021	Anti-Spoofing	Detection	mmWave Radar	real-world data
	Nallabolu et al.[267]	2021	Anti-Spoofing	Detection	mmWave Radar	real-world data
	SAS[284]	2010	Anti-Spoofing	Authentication	GNSS	Simulation data
	NMA[97]	2016	Anti-Spoofing	Authentication	GNSS	Real-world data, Simulation data
	Chimera[10]	2017	Anti-Spoofing	Authentication	GNSS	Real-world data, Simulation data
	Wang et al.[351]	2017	Anti-Spoofing	Mitigation	GNSS	Simulation data, TEXBAT
	Eldosouky et al.[81]	2019	Anti-Spoofing	Mitigation	GNSS	Simulation data
	Zhou et al.[478]	2023	Anti-Spoofing	Mitigation	GNSS	TEXBAT
	Hong et al.[129]	2022	Anti-Spoofing	Mitigation	IMU	Simulation data
	UNROCKER[152]	2023	Anti-Spoofing	Mitigation	IMU	Real-world data, Simulation data
	VIMU[375]	2024	Anti-Spoofing	Mitigation	IMU	Real-world data, Simulation data
	Zhang et al.[442]	2020	Anti-Spoofing	Mitigation	mmWave Radar	real-world data
	Chen et al. [48]	2025	Anti-Spoofing	Mitigation	Camera/ADAS	openpilot, CARLA
	Swinney et al.[321]	2021	Anti-Jamming	Detection	GNSS	Custom dataset
	Spanghero et al.[309]	2025	Anti-Jamming	Detection	GNSS	Jammertest, Real-world data
	Wang et al.[348]	2021	Anti-Jamming	Mitigation	GNSS	Simulation data
	MFMC[149]	2022	Anti-Jamming	Mitigation	GNSS	Custom dataset

Acoustic injection exploits mechanical resonance in inertial sensors by emitting ultrasonic or audible signals. For IMUs, Son et al. [306] showed that single-tone acoustic excitation can induce gyroscope deviations. WALNUT [338] combines acoustic induction with ADC aliasing to trigger false readings, and KITE [105] refines resonance-response modeling for more precise control. For ultrasonic ranging systems, Yan et al. [406], Xu et al. [403], and Gluck et al. [108] demonstrated that crafted acoustic echoes can be injected to spoof distance measurements.

RF injection targets radio-frequency sensors such as mmWave radar by introducing synchronized or tailored electromagnetic signals. Sun et al. [320] and Komissarov et al. [179] injected synchronized RF signals to manipulate radar point clouds, while Vennam et al. [343] synthesized customized spoofing waveforms to generate deceptive objects.

Physical manipulation involves placing adversarial objects or engineered surfaces in the environment to passively spoof sensors. For mmWave radar, MetaWave [57], TileMask [484], and mmHide [106] employ metamaterial surface patterns to control radar reflections.

Jamming. Jamming attacks disrupt or block legitimate sensor signals through interference, thereby degrading or completely disabling sensor functionality. These attacks exploit physical-layer vulnerabilities by overwhelming or corrupting the sensing modality with high-energy or carefully crafted noise across acoustic, electromagnetic, or optical domains.

Acoustic interference employs high-intensity sound to jam ultrasonic sensors. For ultrasonic ranging, Lim et al. [217] demonstrated that high-power acoustic jamming can cause missed detections and destabilize autonomous control policies. Electromagnetic interference disrupts sensor operation through radiated or conducted RF noise. For IMUs, Jang et al. [150] presented a remote electromagnetic injection attack that corrupts communication between the IMU and flight controller, leading to system failure.

2.4.2 Sensor Defenses

Defenses for motion perception apply anti-spoofing and anti-jamming mechanisms to detect and recover from corrupted sensor signals.

Anti-Spoofing Defenses. Anti-spoofing defenses detect, authenticate, and mitigate forged signals through three complementary strategies: Detection, Authentication, and Mitigation. Detection identifies malicious signals or anomalous sensor readings by analyzing inconsistencies in signal characteristics, sensor outputs, or cross-modal correlations.

For GNSS, Falco et al. [90] used dual-antenna double-difference dispersion to detect spatially inconsistent signals. Crowd-GPS-Sec [151] analyzes temporal and spatial inconsistencies across a crowd of devices to identify outliers. DeepSIM [405] uses Siamese networks to match ground-level imagery with satellite maps for consistency verification. DeepPOSE [161] reconstructs vehicle speed and trajectory using ConvLSTM to detect implausible motion patterns. Iqbal et al. [147] introduced a VAE-WGAN framework for zero-day spoofing detection. PADS [232] fuses GNSS with Wi-Fi and cellular data to improve robustness. Jin et al. [165] employed anti-jamming antenna arrays coupled with LightGBM for real-time spoofing classification.

For IMU, SDI [329] introduces cross-sensor consistency checks, Liu et al. [231] localized acoustic sources via MLPs, and CPD-MhIMU [295] deploys heterogeneous IMUs with adaptive EKF fusion. For ultrasonic ranging, Xu et al. [403] introduced physical shift authentication. SoundFence [240] randomizes pulse periods, and SecureTrack [304] incorporates EMI monitoring. For mmWave radar, Sun et al. [320] proposed challenge-response mechanisms, and Nallabolu et al. [267] designed hybrid-slope chirps.

Authentication cryptographically verifies the authenticity and integrity of sensor signals to ensure they originate from legitimate sources. For GNSS, SAS [284] introduces encrypted authentication sequences to bind signals to their true source. NMA [97] demonstrates navigation message authentication for Galileo based on the TESLA protocol. Chimera [10] proposes time-binding tags that link signal transmission to precise time slots, preventing replay.

Table 10 A summary of **adversarial** attacks and defenses for **cross-modal perception**.

Attack/Defe	Method	Year	Category	Subcategory	Target Model	Dataset
Adversarial Attack	Li et al.[189]	2023	Digital Attack	Point Injection	Fusion 3D Detector	nuScenes
	DejaVu[133]	2025	Digital Attack	Temporal Misalignment	Fusion 3D Detector	nuScenes
	Cao et al.[42]	2021	Physical Attack	Adversarial Object	Fusion 3D Detector	KITTI, nuScenes
	Hallyburton et al.[118]	2022	Physical Attack	LiDAR Spoofing	Fusion 3D Detector	KITTI, nuScenes
	Li et al.[207]	2024	Physical Attack	Adversarial Object	Fusion 3D Detector	nuScenes
Adversarial Defense	Wang et al.[361]	2022	Robust Training	Adversarial Training	Fusion 3D Detector	KITTI
	MMCert[373]	2024	Robust Inference	Certified Defense	Multi-Modal Model	Kinetics-400, Food-101
	Yang et al.[411]	2024	Robust Inference	Input Sanitization	Fusion 3D Detector	nuScenes, KITTI

Mitigation actively counteracts or reduces attack effects. For GNSS, Wang et al. [351] introduced MLE-based localization and cancellation, Eldosouky et al. [81] used cross-UAV localization, and Zhou et al. [478] proposed VTL-based correction pipelines. For IMU, Hong et al.'s work [129] combines LSTM prediction with CUSUM monitoring, UNROCKER [152] applies denoising autoencoders, and VIMU [375] integrates physical modeling with anomaly detection. For mmWave radar, Zhang et al. [442] introduced VANET-based coordination. For camera-based ADAS, Chen et al. [48] evaluated automated and human-driver safety interventions in open-source ADAS against adversarial patch attacks, analyzing intervention conflicts and their resolution to enhance system resilience.

Anti-Jamming Defenses. Anti-jamming defenses enhance receiver resilience against intentional or unintentional interference through two primary strategies: Detection, which identifies jamming signals or interference patterns, and Mitigation, which suppresses or filters interference to restore signal integrity and functionality.

For detection of jamming in GNSS, Swinney et al. [321] fused frequency- and time-domain representations using VGG16 with transfer learning to detect jamming. Spanghero et al. [309] leveraged VTOL UAVs to localize jammers by analyzing spatial signal degradation patterns.

For mitigation of jamming in GNSS, Wang et al. [348] demonstrated that reservoir computing and LSTM networks can reconstruct GPS signals corrupted by jamming. MFMC [149] develops multi-frequency, multi-constellation receivers to exploit signal diversity and reduce vulnerability.

2.5 Cross-Modal Perception

Modern embodied systems increasingly rely on multi-sensor fusion (combining cameras, LiDAR, and radar) to achieve robust perception. Cross-modal perception introduces unique vulnerabilities absent from single-modality systems: attackers can exploit inconsistencies between modalities, corrupt fusion mechanisms, or target the weakest channel to compromise the entire perception pipeline. These attacks are particularly dangerous because they can bypass defenses designed for individual modalities. This subsection reviews vulnerabilities and defenses in multi-modal perception models, organized into two parts: **Adversarial Attacks** (Section 2.5.1) target sensor fusion pipelines through cross-channel perturbations and temporal misalignment; and **Adversarial Defenses** (Section 2.5.2) protect fusion systems through certified robustness, adversarial training, and modality-specific sanitization.

2.5.1 Adversarial Attacks

Multi-sensor fusion combines complementary modalities to improve perception accuracy and robustness, but the fusion process itself introduces attack surfaces, particularly at the feature alignment, projection, and decision stages.

Digital Attacks. Li et al. [189] showed that fusion models can be deceived by manipulating only the LiDAR channel, achieving a 99% attack success rate with as few as 200 adversarial points injected into the point cloud. DejaVu [133] exploits synchronization dependencies between sensors: a single-frame LiDAR delay causes 88.5% mAP degradation in 3D detection, while three-frame camera delays reduce MOT performance by 73% MOTA, revealing that temporal misalignment is a critical but underprotected attack surface.

Physical Attacks. Physical attacks on cross-modal perception exploit the geometric correspondence between 2D images and 3D point clouds. Cao et al. [42] presented the first physical-world attack that simultaneously fools both camera and LiDAR channels using 3D-printed adversarial objects, demonstrating that physically realizable perturbations can evade all tested fusion-based detectors. Hallyburton et al. [118] proposed the frustum attack that compromises all eight widely-used perception algorithms (both LiDAR-only and camera-LiDAR fusion) through black-box LiDAR spoofing, while remaining stealthy to existing defenses. Li et al. [207] conducted the first comprehensive study attacking all three sensing modalities simultaneously (camera, LiDAR, and radar) using a single adversarial object that exploits cross-modal geometric constraints.

2.5.2 Adversarial Defenses

Defenses for cross-modal perception systems exploit redundancy between modalities and enforce consistency constraints to detect or mitigate multi-channel attacks.

Robust Training. Adversarial training for fusion models must account for cross-channel externalities. Wang et al. [361] discovered that single-channel adversarial training in fusion models can reduce robustness to attacks on other channels (a cross-channel externality) and propose multi-channel adversarial training as a countermeasure.

Robust Inference. Wang et al. [373] proposed MMCert, the first certified defense against adversarial attacks on multi-modal models, deriving provable robustness bounds through randomized smoothing across modalities. Yang et al. [411] developed a robust multi-sensor fusion model that applies modality-specific input sanitization before fusion, preventing adversarial patches on one modality from corrupting the joint representation.

3 Cognition

Cognition forms the second layer, encompassing perception while adding semantic interpretation and logical inference, expanding the agent's capability from sensing to understanding. This expansion introduces new attack surfaces beyond perceptual corruption: adversaries can now manipulate how agents interpret instructions, exploit world model hallucinations, and hijack reasoning chains. With LLMs and VLMs increasingly serving as the "brain" of embodied systems, the cognitive layer becomes both the engine of intelligent behavior and a high-value target for adversarial manipulation. Cognitive threats differ fundamentally depending on which function they target (language comprehension, world simulation, and multi-step reasoning), each demanding distinct adversarial analyses. This section organizes cognition by targeted function: **Instruction Understanding** (Section 3.1) addresses jailbreak attacks that manipulate natural language instructions to bypass safety constraints and induce harmful physical actions, along with corresponding defenses and benchmarks; **World Model** (Section 3.2) examines hallucination in scene understanding and rule violations in predictive models; and **Reasoning** (Section 3.3) covers chain-of-thought hijacking attacks that corrupt deliberative reasoning.

Table 11 A summary of **cognitive attacks and defenses** for **embodied cognition**.

Attack/Defe	Method	Year	Category	Subcategory	Target Model	Dataset/Benchmark
Cognitive Attack	CHAI [37]	2024	Instruction Attack	Jailbreak Attacks	Embodied LLM	Simulation, Real World
	BadNAVer [22]	2025	Instruction Attack	Jailbreak Attacks	Navigation Agent	Matterport3D
	Chen et al. [58]	2024	World Model tack	At-Hallucination	VLM	Custom
	Tao et al. [326]	2025	World Model tack	At-Hallucination	VLM	Custom
	Baraldi et al. [24]	2025	World Model tack	At-Hallucination	World Model	Custom
	MASH-VLM [132]	2025	World Model tack	At-Hallucination	Video-LLM	Custom
	HRSSM [313]	2024	World Model tack	At-Rule Violation	World Model	Custom
	Wen et al. [386]	2024	World Model tack	At-Rule Violation	World Model	Custom
	SafeDreamer [141]	2024	World Model tack	At-Rule Violation	World Model	Safety Gymnasium
	Drive-WM [377]	2024	World Model tack	At-Rule Violation	World Model	nuScenes
	Li et al. [205]	2025	World Model tack	At-Rule Violation	World Model	Custom
	VL-SAFE [287]	2025	World Model tack	At-Rule Violation	World Model	AD Simulation
	Chakraborty et al. [45]	2025	Reasoning Attack	Grounding Failure	Embodied LLM	ALFWorld
	Han et al. [119]	2026	Reasoning Attack	Grounding Failure	Embodied LLM	Custom
	H-CoT [181]	2025	Reasoning Attack	CoT Hijacking	Reasoning Model	Custom
Cognitive Defense	J-DAPT [**?**]	2025	Instruction Defense	Jailbreak Defenses	Embodied LLM	nuScenes, Maritime, Quadruped
	Abuduweili et al. [3]	2025	Instruction Defense	Jailbreak Defenses	Embodied LLM	Gazebo, Real World

3.1 Instruction Understanding

Embodied agents rely on natural language instructions to bridge human intent and physical action. Unlike text-only chatbots, where misinterpretation produces merely incorrect text, failures in embodied instruction understanding can trigger unsafe physical behaviors such as collisions, property damage, or human injury. This subsection surveys jailbreak attacks on embodied instruction following, corresponding defenses, and benchmarks for evaluating instruction safety.

Jailbreak Attacks. Jailbreak attacks manipulate language inputs to circumvent safety constraints, causing embodied agents to execute harmful physical actions that would normally be refused. CHAI [37] optimizes adversarial commands against the LVLM command layer of physical agents in the white-box setting. Black-box attacks at the sentence level craft semantically meaningful prompts that exploit weaknesses in safety filters. BadNAVer [22] demonstrates that jailbreaks in embodied navigation directly trigger unsafe physical actions.

Jailbreak Defenses. Defenses against jailbreaks aim to detect or block adversarial instructions before they are translated into physical actions. J-DAPT [260] introduces multimodal domain adaptation for robotic jailbreak detection, using vision-language alignment to identify adversarial instructions before they reach the control pipeline. Abuduweili et al. [3] integrated reachability analysis with LLM-controlled robots, providing formal safety guarantees by rejecting instructions whose predicted outcomes violate verified safety envelopes.

Benchmarks. IndustryEQA [203] extends embodied QA to safety-critical industrial environments with hazard recognition and compliance verification. SQA3D [420] introduces situated question answering in 3D scenes requiring spatial and commonsense reasoning. MMRO [195] benchmarks multimodal LLMs as cognitive engines for in-home robotics, revealing safety as a persistent weakness. MetaVQA [356] fine-tunes VLMs with embodied scene data to improve spatial reasoning in safety-critical driving simulations. AGENTSAFE [220] evaluates embodied agent vulnerability to jailbreaks across adversarial scenarios with hazardous tasks. EmbodiedBench [410] provides a comprehensive evaluation framework with explicit safety metrics.

3.2 World Model

World models enable embodied agents to predict future states, reason about physical dynamics, and evaluate action consequences before execution. When these internal representations diverge from physical reality (through hallucination, sim-to-real gaps, or prediction failures), agents make decisions based on false beliefs about their environment, with potentially catastrophic physical consequences. This subsection surveys threats to world model safety organized by two failure modes: **Hallucination in Scene Understanding** addresses VLM and world model hallucination that generates nonexistent objects, spatial relations, or actions; and **Rule Violation** covers predictive model failures and emergent misalignment that cause agents to violate physical laws, domain-specific rules, or safety constraints.

Hallucination in Scene Understanding. VLM hallucination, i.e., generating descriptions of objects, spatial relations, or actions that do not exist in the physical scene, poses acute risks when these models serve as the perceptual backbone of embodied agents. Chen et al. [58] showed that multi-object hallucination in VLMs remains pervasive in embodied scene understanding, and Tao et al. [326] demonstrated that hallucination is especially severe in visual-text tasks for embodied agents. MASH-VLM [132] disentangles spatial and temporal tokens via DST-attention to reduce action-scene misattribution in video-LLMs. Beyond VLM hallucination, world models used for internal simulation exhibit distinct pathologies. Baraldi et al. [24] identified scene-generation pathology criteria spanning temporal consistency, physical conformity, and condition consistency, confirming a systematic safety gap in current world model predictions.

Rule Violation. Beyond hallucination, world models can systematically violate physical laws, domain-specific rules, and safety constraints; such failures directly translate into unsafe embodied behavior. Predictive model failures occur when learned dynamics models compound errors over long horizons, producing increasingly dangerous predictions. Li et al. [205] surveyed world model architectures across RSSM, Transformer, diffusion, and other paradigms, identifying error accumulation, distribution shift, and physical

consistency as critical safety challenges. HRSSM [313] learns latent dynamic robust representations to improve world model resilience to distribution shift. Wen et al. [386] found that compounding errors in video prediction rollouts limit long-horizon reliability. Safety-aware world models explicitly incorporate constraints into the prediction-planning loop: SafeDreamer [141] integrates Lagrangian-based safety constraints into the Dreamer framework, VL-SAFE [287] supervises world models using VLM-derived safety scores for autonomous driving, and Drive-WM [377] uses multi-view diffusion for safer trajectory selection, though training planners on WM-generated data creates a cascading risk where pathologies propagate to downstream policies.

3.3 Reasoning

Reasoning addresses vulnerabilities in the deliberative processes that embodied agents use for multi-step problem solving.

Embodied reasoning has evolved rapidly since 2022, when large language models were first shown to decompose high-level instructions into grounded action plans. SayCan [6] introduced affordance-based grounding, scoring LLM-proposed actions by physical feasibility, while Inner Monologue [142] closed the loop by incorporating environment feedback into the reasoning process. Subsequent work scaled these ideas to end-to-end multimodal architectures: RT-2 [33] co-fine-tuned vision-language models on robot data and exhibited emergent semantic reasoning (e.g., selecting an improvised tool), and π_0 [30] extended the paradigm to flow-matching VLAs capable of complex dexterous tasks. Most recently, embodied chain-of-thought methods such as ECoT [430] and CoT-VLA [464] train VLAs to produce explicit intermediate reasoning traces before acting, improving task success by over 28%. As these reasoning capabilities grow more powerful, they simultaneously enlarge the attack surface: errors or adversarial manipulations at the reasoning stage can cascade into unsafe physical actions, making the security of embodied reasoning an increasingly urgent concern.

Grounding Failure. A complementary failure mode arises when reasoning is logically coherent but physically infeasible. Chakraborty et al. [45] showed that scene-task inconsistencies increase embodied agent hallucination rates by up to 40×, with models repurposing available objects rather than rejecting infeasible instructions. Han et al. [119] further demonstrated that LLMs instructed to navigate during a simulated fire drill directed a robot toward a server room instead of the emergency exit, revealing that high overall accuracy can mask critical grounding failures in safety-sensitive contexts.

Chain-of-Thought Hijacking. Chain-of-thought (CoT) reasoning enables transparent multi-step deliberation but exposes intermediate reasoning steps to adversarial manipulation. H-CoT [181] demonstrates that inserting adversarial steps into chain-of-thought traces can hijack reasoning models toward harmful conclusions.

4 Planning

Planning forms the third layer, encompassing perception and cognition while adding the generation of action sequences, expanding the agent's capability from understanding to decision-making. This expanded capability extends the attack surface from passive interpretation to active goal pursuit: adversaries can now corrupt task decomposition, hijack trajectory optimization, and manipulate multi-agent coordination strategies. Modern embodied planners increasingly leverage LLMs for high-level task decomposition while relying on traditional methods for low-level motion generation. Because planning operates at multiple levels of abstraction, from symbolic task decomposition through continuous trajectory optimization to distributed multi-agent coordination, each level presents qualitatively different vulnerabilities and demands distinct security considerations. This section organizes planning into three subsections: **Task Planning** (Section 4.1) addresses vulnerabilities in LLM-based task decomposition, chain-of-thought reasoning, and goal specification, including jailbreak attacks that manipulate planners into generating harmful action sequences, as well as goal hijacking and reward hacking; **Trajectory Planning** (Section 4.2) covers threats to trajectory prediction and path planning, including adversarial perturbation of collision avoidance systems; and

Multi-Agent Planning (Section 4.3) discusses planning-time coordination challenges including distributed task allocation, consensus failures, subgoal manipulation, and goal conflicts among cooperative agents. Note that multi-agent planning focuses on the **planning phase** (who does what), while execution-time collaboration is covered in Section 5.

4.1 Task Planning

Task planning translates high-level objectives into actionable subgoal sequences, increasingly through LLM-based decomposition, chain-of-thought reasoning, and tool use. These capabilities enable flexible, generalizable planning but also introduce novel attack surfaces: adversaries can manipulate task specifications to induce unsafe decompositions, hijack reasoning chains to subvert intended goals, or exploit jailbreak vulnerabilities to elicit harmful plans. This subsection examines adversarial attacks on classical optimization-based planners and modern LLM-based task decomposition, jailbreak attacks that manipulate planners into generating harmful action sequences, backdoor attacks that implant hidden triggers, jailbreak defenses that prevent malicious instruction injection, and emerging risks including unforced constraints violation.

4.1.1 Adversarial Attacks

Adversarial attacks on task planners primarily target black-box threat models where attackers perturb inputs or environmental states without model access. Islam et al. [148] demonstrated that small visual perturbations can mislead CLIP-based vision–language navigation systems into attacker-defined paths, highlighting vulnerabilities in vision-grounded task planners. Vemprala and Kapoor [342] showed that adversarial state configurations can degrade eigenstructure in classical optimization-based planners, forcing failure or excessive computation.

4.1.2 Jailbreak Attacks

Jailbreak attacks manipulate LLM-based planners by crafting inputs that bypass safety guardrails to elicit harmful task decompositions.

White-box Attacks. White-box attacks leverage gradient-based optimization to craft adversarial suffixes or token-level perturbations. EIRAD [230] adapts gradient-based suffix optimization, appending adversarial tokens to benign inputs so that untargeted variants divert the agent from the intended task while targeted variants steer outputs toward specific harmful goals. POEX [244] improves suffix quality through a mutator–selector–evaluator loop that balances jailbreak success with action executability, validating attacks on real robots and introducing Harmful-RLBench, a paired benign–harmful task suite for sim-to-real safety evaluation.

Black-box Attacks. Black-box attacks manipulate instructions without model access, operating through prompt engineering and semantic manipulation. RoboPAIR [292] automates jailbreak generation by extending PAIR with robot-specific prompts and a syntax checker that enforces API-compliant action formats. BADROBOT [437] identifies three embodiment-specific attack surfaces: contextual jailbreak, safety misalignment, and conceptual deception, showing that LLM-driven robots can be coerced into unsafe actions using in-the-wild prompts across both simulation and physical platforms.

4.1.3 Backdoor Attacks

Backdoor attacks implant hidden triggers into task planners during pre-training or fine-tuning while preserving clean-task performance, enabling malicious behaviors to activate only when the trigger appears at deployment. CBA [219] poisons in-context demonstrations using combined textual and visual triggers to activate harmful behaviors at deployment. BALD [164] taxonomizes backdoor pathways in LLM-based planners, covering word-level triggers, scenario manipulation, and RAG-based knowledge injection. RoboTroj [265] demonstrates backdoor attacks via poisonous fine-tuning of soft-prompts to inject malicious plans when trigger words appear in task descriptions.

Table 12 A summary of attacks and defenses for **embodied planning**.

Attack/Defense	Method	Year	Category	Subcategory	Target Model	Environment
Adversarial Attack	Zhang et al. [443]	2022	White-box	Optim.-Based Attack	Trajectory Planner	Apolloscape, NGSIM, nuScenes
	AdvDO [43]	2022	White-box	Optim.-Based Attack	Trajectory Planner	nuScenes
	KING [123]	2022	White-box	Optim.-Based Attack	Trajectory Planner	CARLA
	Adv-GAN [91]	2024	White-box	Model-Based Attack	Trajectory Planner	Apolloscape, NGSIM, nuScenes
	ADvLM [447]	2024	White-box	Optim.-Based Attack	Trajectory Planner	nuScenes, DriveLM
	Islam et al. [148]	2024	Black-box	Optim.-Based Attack	Task Planner	EnvLarge-10
	AdvSim [357]	2021	Black-box	Optim.-Based Attack	Trajectory Planner	UrbanScenarios
	STRIVE [289]	2022	Black-box	Optim.-Based Attack	Trajectory Planner	nuScenes
	Zheng et al. [472]	2023	Black-box	Optim.-Based Attack	Trajectory Planner	nuScenes, Argoverse
	UTCIA [21]	2025	Black-box	Optim.-Based Attack	Trajectory Planner	Trajectory Datasets
	Avatar [225]	2025	Black-box	Optim.-Based Attack	Trajectory Planner	Waymax
	LC [77]	2020	Black-box	Model-Based Attack	Trajectory Planner	CARLA
	AdvDiffuser [396]	2024	Black-box	Model-Based Attack	Trajectory Planner	nuScenes
	NADE [96]	2021	Black-box	Model-Based Attack	Trajectory Planner	CARLA
	Szvoren et al. [322]	2025	Black-box	Physical Attack	Trajectory Planner	Gazebo, Unitree Go1
Adversarial Defense	Thumm et al. [330]	2023	Robust Training	Safety Constraints	Trajectory Planner	OpenAI safety gym
	AR-ICRL [402]	2024	Robust Training	Safety Constraints	Trajectory Planner	Blocked Half-Cheetah/Ant/Walker
	SMPC [35]	2021	Robust Inference	Output Moderation	Trajectory Planner	Matlab
	Yurtsever et al. [429]	2019	Robust Inference	Output Moderation	Trajectory Planner	NuDrive
	CSP-GAN-LSTM [262]	2023	Robust Inference	Output Moderation	Trajectory Planner	NGSIM, highD
Jailbreak Attack	EIRAD [230]	2024	White-box	Word-Level	Task Planner	AI2-THOR
	POEX [244]	2025	White-box	Word-Level	Task Planner	CoppeliaSim, RLBench, Real world
	RoboPAIR [292]	2024	Black-box	Sentence-Level	Task Planner	nuScenes, Real world
	BADROBOT [437]	2025	Black-box	Sentence-Level	Task Planner	RLBench, Real world
	Zhang et al. [449]	2024	Black-box	Sentence-Level	Trajectory Planner	EyeSim VR
	Wen et al. [383]	2024	Black-box	Sentence-Level	Trajectory Planner	Touchdown, Map2seq
	PINA [226]	2026	Black-box	Sentence-Level	Trajectory Planner	Indoor/Outdoor Navigation
Jailbreak Defense	SafeEmbodAI [448]	2024	Robust Inference	Safe Prompt	Trajectory Planner	EyeSim VR
	NPE [383]	2024	Robust Inference	Safe Prompt	Trajectory Planner	Touchdown, Map2seq
	J-DAPT [260]	2025	Robust Inference	Jailbreak Detection	Trajectory Planner	nuScenes, Maritime, Quadruped
	RoboSafe [359]	2025	Robust Inference	Runtime Safety	Task/Traj Planner	AI2-THOR, MetaWorld
	CEE [407]	2025	Robust Inference	Representation Eng.	Task Planner	EI Safety Benchmarks
	SafePlan [275]	2025	Robust Inference	Safe Prompt	Task Planner	AI2-THOR, Synthetic
	Zhang et al. [450]	2025	Robust Inference	Runtime Safety	Trajectory Planner	EyeSim, Real world
Backdoor Attack	CBA [219]	2024	Data Poisoning	Multi-Modal Triggers	Task Planner	ProgPrompt, VoxPoser, VisProg, Real
	BALD [164]	2025	Training Manipulation	Multi-Modal Triggers	Task Planner	HighwayEnv, CARLA, nuScenes
	Robo-Troj [265]	2025	Training Manipulation	Word-Level Triggers	Task Planner	VirtualHome, AI2-THOR

4.1.4 Jailbreak Defenses

Jailbreak defenses for task planning focus on preventing malicious instruction injection and ensuring that LLM-based planners adhere to safety constraints during deployment. Current strategies employ safe prompting, jailbreak detection, runtime validation, and representation engineering techniques.

SafeEmbodAI [448] integrates safe prompting, state management, and safety validation modules to verify and sanitize actions before execution, preventing unsafe navigation behaviors. NPE [383] employs structured templates such as Chain-of-Thought and Plan-and-Solve to improve planner robustness against text-based manipulations. J-DAPT [260] integrates textual and visual embeddings via attention-based fusion and adapts general jailbreak datasets to robotics-specific domains for multimodal detection. RoboSafe [359] combines backward reflective reasoning over recent trajectories with forward predictive reasoning from safety memory to generate executable predicate-based safety logic. Concept Enhancement Engineering (CEE) [407] steers internal representations toward safe concepts to defend against jailbreak attacks in embodied AI systems. SafePlan [275] interposes formal logic verification at multiple points in the CoT pipeline to filter unsafe robotic task plans. A unified framework for security and safety in LLM-integrated robotic systems [450] combines interpretable prompting, state-aware planning, and real-time validation to jointly address safety and prompt-injection security in LLM-driven mobile robots.

4.2 Trajectory Planning

Trajectory planning generates continuous trajectories that satisfy kinematic, dynamic, and safety constraints. Modern approaches combine learned trajectory prediction with classical path planning and collision avoidance, creating hybrid systems that inherit vulnerabilities from both paradigms. Attacks on trajectory planners can induce collisions, amplify prediction errors, or degrade trajectory quality through adversarial perturbations to perception inputs, prediction models, or planning algorithms. This subsection examines adversarial attacks on trajectory prediction and path planning, jailbreak attacks on LLM-based navigation planners, and adversarial defenses through robust training and inference.

4.2.1 Adversarial Attacks

Adversarial attacks on trajectory planners exploit weaknesses in trajectory prediction and generation models, driving agents toward unsafe maneuvers or systematically amplifying prediction errors. These attacks fall into two main classes: white-box attacks, which use gradient-based methods to craft perturbations on trajectories or map context with full model access, and black-box attacks, which leverage query-based optimization or generative models to synthesize realistic adversarial scenarios without model access.

White-box Attacks. White-box attacks leverage model gradients to craft precise adversarial perturbations on trajectories or context maps. Zhang et al. [443] perturbed nominal vehicle trajectories to maximize prediction error in trajectory forecasting models. AdvDO [43] uses a differentiable dynamics model to construct plausible adversarial trajectories that mislead downstream planners. KING [123] employs a differentiable kinematic model to efficiently search for critical but feasible scenes. Adv-GAN [91] employs an LSTM-based generator to produce adversarial trajectory perturbations and refines them using model predictive control under realism and safety constraints. ADvLM [447] addresses textual instruction variability and time-series visual scenarios in VLM-based autonomous driving through Semantic-Invariant Induction for diverse prompt libraries and Scenario-Associated Enhancement for frame-perspective optimization.

Black-box Attacks. Black-box attacks operate without model access, relying on query-based optimization, surrogate models, or RL to generate failure-inducing scenarios. AdvSim [357] provides a general black-box adversarial scenario search framework for end-to-end planners. STRIVE [289] perturbs real-world scenes in the latent space of a VAE-based traffic motion model to generate challenging scenarios for stress-testing. Zheng et al. [472] introduced adversarial corruption of context maps required by trajectory predictors. UTCIA [21] generates universal black-box adversarial perturbations for trajectory representation learning. Avatar [225] uses RL to optimize adversarial trajectories without model access. LC [77] trains RL agents to act

as adversaries, intentionally inducing collisions and exposing weaknesses in planners. AdvDiffuser [396] leverages diffusion guidance to synthesize realistic yet failure-inducing trajectories. NADE [96] generates naturalistic adversarial driving scenarios to stress-test end-to-end planners.

Physical adversarial attacks on robot trajectory planners [394] characterize how manipulating the environment can cause trajectory planner failures in real-world deployments. JackZebra [316] demonstrates long-horizon goal hijacking through adversarial patches on an attacker vehicle, gradually steering a victim AV to an attacker-chosen destination.

4.2.2 Jailbreak Attacks

Jailbreak attacks targeting LLM-based navigation systems manipulate natural language instructions to bypass safety constraints and elicit unsafe navigation behaviors. Unlike task planning jailbreaks that corrupt high-level goal decomposition, navigation jailbreaks directly compromise low-level trajectory generation.

Zhang et al. [449] modeled Obvious Malicious Injection (OMI) and Goal Hijacking Injection (GHI) against LLM-integrated mobile robots. Wen et al. [383] demonstrated that insertion and swap attacks significantly degrade the performance of GPT-3, GPT-4, and LLaMA-based navigation planners on Touchdown and Map2Seq, with errors concentrated at intersections and other high-ambiguity locations. PINA [226] extends prompt injection to directly misguide physical navigation, leading to unsafe routes and mission failure.

4.2.3 Adversarial Defenses

Adversarial defenses for trajectory planning operate through robust training, which incorporates safety constraints during learning, and robust inference, which applies output moderation at deployment.

Robust Training. Robust training integrates safety constraints into the learning process. Thumm et al. [330] proposed proactive replacement and projection methods that modify agent actions during RL training to reduce failsafe interventions, yielding policies with fewer safety violations. AR-ICRL [402] extends inverse RL to infer safety constraints from expert demonstrations that remain valid even under model misspecification.

Robust Inference. Robust inference defends planners at deployment through output moderation. SMPC [35] uses stochastic model predictive control with backup trajectories computed via reachable sets, overwriting planned outputs when safety constraints are violated. Yurtsever et al. [429] identified hazardous behaviors at runtime, enabling planners to filter or down-weight high-risk maneuvers. CSP-GAN-LSTM [262] combines convolutional pooling with attention-based trajectory prediction to compute collision risk via time-to-collision metrics during inference.

4.3 Multi-Agent Planning

Multi-agent planning extends single-agent task and trajectory planning to teams of embodied agents that must jointly decompose tasks, allocate subtasks, and synthesize coordinated plans. This distributed setting introduces unique attack surfaces: an adversary can compromise a single agent's planner to inject malicious subtasks that cascade through the team, manipulate inter-agent communication to corrupt plan consensus, or exploit Byzantine faults to subvert collective decision-making.

4.3.1 Byzantine Faults

Byzantine faults arise when agents exhibit arbitrary or malicious behavior during distributed planning, corrupting consensus formation and task allocation. Strobel and Ferrer [310] demonstrated that classical consensus algorithms break down under Byzantine attacks in swarm robotics. Blumenkamp and Prorok [32] showed that self-interested agents in multi-robot planning tasks learn manipulative communication strategies through a differentiable shared channel, suggesting that adversarial behavior may emerge naturally from competitive pressure. He et al. [126] introduced the Agent-in-the-Middle (AiTM) attack that intercepts and manipulates messages between LLM-based agents during cooperative planning. Schroeder de Witt [74]

Table 13 A summary of emerging risks and defenses for **multi-agent planning**.

Attack/Defense	Method	Year	Category	Subcategory	Target Model	Environment
Emerging Risks	Strobel and Fer-rer [310]	2020	Multi-Agent	Byzantine Faults	Multi-Agent Planner	Physical Robots
	Blumenkamp et al. [32]	2021	Multi-Agent	Byzantine Faults	Multi-Agent Planner	Coverage, Path Planning
	He et al. [126]	2025	Multi-Agent	Byzantine Faults	Multi-Agent Planner	Multi-Agent Frameworks
	Choudhury et al. [70]	2022	Multi-Agent	Goal Conflicts	Multi-Agent Planner	Task Allocation
	Khamis et al. [172]	2024	Multi-Agent	Goal Conflicts	Multi-Agent Planner	Task Allocation
	Bahrami and Jafarne-jadsani [18]	2025	Multi-Agent	Goal Conflicts	Multi-Agent Planner	Relative Localization
	Li et al. [194]	2020	Multi-Agent	Potential Defenses	Multi-Agent Planner	Multi-Robot Networks
	Strobel et al. [311]	2023	Multi-Agent	Potential Defenses	Multi-Agent Planner	24 Physical Robots
	Lee and Panagou [184]	2025	Multi-Agent	Potential Defenses	Multi-Agent Planner	Distributed Control
	Gandhi et al. [104]	2025	Multi-Agent	Potential Defenses	Multi-Agent Planner	Physical Robots

taxonomizes multi-agent security threats including cascading failures, monoculture collapse, and conformity bias that drives false consensus on unsafe plans.

4.3.2 Goal Conflicts

Adversarial or self-interested agents exploit cooperative planning protocols to advance conflicting objectives. Choudhury et al. [70] formulated robust task allocation strategies that maintain plan quality under adversarial cost perturbation. Khamis et al. [172] identified security-relevant gaps in multi-robot task allocation including lack of authentication and absence of Byzantine robustness guarantees. Bahrami and Jafarnejadsani [18] examined how adversarial perception attacks propagate through multi-robot relative localization to corrupt downstream coordination and planning. Zhou and Tokekar [475] reviewed algorithmic trends for robust multi-robot coordination under adversarial agents, while Sookha and Benevenuto [308] provided a taxonomy of adversarial attacks on multi-agent reinforcement learning that can corrupt learned planning policies.

4.3.3 Potential Defenses

Resilient algorithms ensure that cooperative planning converges correctly despite misbehaving agents. Li et al. [194] proposed a centerpoint-based aggregation rule that guarantees convergence to the true target state even when adversarial robots inject arbitrary state estimates. Strobel et al. [311] deployed smart contracts that regulate a crypto-token economy among physical robots, causing Byzantine robots to exhaust their tokens and be neutralized. Lee and Panagou [184] designed a CBF-based distributed controller that guarantees resilient consensus and collision avoidance using only locally available information. Gandhi et al. [104] presented RoboRebound, extending Byzantine fault tolerance to physical multi-robot systems where adversarial agents can block paths or cause collisions.

4.4 Benchmarks

Simulation platforms, scenario-generation tools, and benchmarks constitute the foundational infrastructure for developing and evaluating embodied planners. These components differ in fidelity, scalability, and safety focus, yet collectively enable systematic stress-testing of algorithms. Simulation environments provide arenas for agent interaction, scenario design frameworks construct complex and safety-critical situations, and benchmarks integrate both into standardized evaluation pipelines. Together, they enable systematic stress-testing of planning algorithms under adversarial, rare, and out-of-distribution conditions.

Simulation Platforms. Simulation platforms for embodied planning vary in realism, efficiency, and task coverage. For autonomous driving, SUMMIT [39] provides high-fidelity 3D towns with diverse agents

and configurable weather, while HIGH-ENV [188] offers a lightweight 2D setup for rapid prototyping. CARLA [80] extends fidelity through configurable vehicles, pedestrians, and scenes, and MetaDrive [199] uses procedural generation to produce large distributions of driving layouts. NAVSIM [73] complements these with dataset-replay environments for cost-effective evaluation. In robotics, platforms emphasize physics accuracy and multi-task support. Gazebo [178] integrates ROS and multiple physics engines for navigation and multi-robot coordination. PyBullet [71] offers an efficient Python API and built-in robot models widely adopted in RL. MuJoCo [334] provides high-precision contact dynamics for locomotion and manipulation. Habitat [285] scales visual navigation in realistic indoor environments, iGibson [192] supports physically grounded manipulation tasks, and NVIDIA Isaac Sim [274] delivers photorealistic rendering with GPU-accelerated physics.

Scenario Design Tools. Scenario design tools build on simulators to specify safety-critical interactions in a structured and repeatable way. CARLA Scenario Runner [80] provides a Python API and OpenSCENARIO support for multi-agent coordination. SCENIC [99] introduces a probabilistic programming language for expressing spatial and temporal relationships, enabling concise specification of rare and complex events. SafeBench [398] integrates eight categories of critical driving scenarios and multiple generation algorithms for systematic safety evaluation. SUMO NETEDIT [239] offers graphical editing of road networks for traffic-scale simulations, and CommonRoad [369] supplies XML-based scenario definitions and a Python API for standardized motion-planning research. Together, these tools span language-based specification, graphical editing, and benchmark-oriented safety testing, enabling reproducible and comprehensive evaluation of embodied planners.

Benchmarks build on simulation and scenario design to create standardized, repeatable pipelines for evaluating embodied planners under adversarial, rare, and out-of-distribution conditions. Bench2Drive [159] offers a closed-loop driving suite with 220 routes spanning diverse weather, traffic, and map settings, isolating core planning skills such as lane keeping, merging, overtaking, and emergency handling. M3Bench [460] targets mobile manipulation with 30,000 pick-and-place tasks across 119 household scenes, providing expert demonstrations and tests of generalization to novel objects and layouts. THOR-EAE [370] assesses both action selection and natural language explanation with 840,000 samples in AI2-THOR. EAI [197] standardizes evaluation for LLM-driven agents, unifying protocols across navigation and interaction, decomposing execution into subgoals, and benchmarking eighteen state-of-the-art models with detailed error analysis.

Safety-focused benchmarks have proliferated to address planning-specific hazards. AgentSafe [220] measures multimodal reasoning in long-horizon navigation with adversarial simulation scenarios and risk-aware task suites inspired by Asimov's Three Laws. HASARD [336] evaluates interactive household manipulation with affordance-level annotations. Safe-BeAI [145] focuses on hazardous and adversarial settings to assess an agent's risk awareness and robustness. SafeAgentBench [419] evaluates embodied agent safety through executable tasks spanning explicit and implicit hazards. AGENTSAFE [259] benchmarks safety of embodied agents on hazardous instructions with multi-stage evaluation across perception, planning, and execution. SafeMindBench [51] benchmarks safety risks in embodied LLM agents.

5 Action and Interaction

Action forms the fourth layer, encompassing perception, cognition, and planning while adding physical execution, expanding the agent's capability from decision-making to real-world interaction. This expansion carries the highest stakes and broadens the attack surface to the physical domain: adversaries can now corrupt control policies to cause collisions, exploit human-agent interaction to endanger people, and poison multi-agent coordination to induce swarm-level failures. In end-to-end models like VLA, this layer represents the full system from visual input to action output. Threats at the action and interaction layer span three distinct scopes of physical engagement (individual robot control, dyadic human-agent encounters, and multi-agent coordination), each escalating in complexity and potential for real-world harm. This section is organized into three subsections: **Robot Control** (Section 5.1) addresses robustness of low-level control policies, including RL-based controllers, diffusion policies, and VLA models; **Human-Agent Interaction**

Table 14 A summary of **adversarial attacks** for **embodied interaction**. For clarity, we append suffixes to the target model (MLP, DP, DT, VLA), where **S**, **A**, **E**, **V**, and **L** denote attack surfaces on State, Action, Environment, Vision, and Language.

Attack	Method	Year	Category	Subcategory	Target Model	Environment
Adversarial Attack	CPA/AA[315]	2020	White-Box	Optim.-Based Attack	MLP-S	Atari, MuJoCo
	RS/MAD[438]	2020	White-Box	Optim.-Based Attack	MLP-S	Atari, MuJoCo
	Weng et al.[387]	2020	White-Box	Optim.-Based Attack	MLP-S	MuJoCo
	MAS/LAS[185]	2020	White-Box	Optim.-Based Attack	MLP-A	Atari, Gym
	DP-Attacker[60]	2024	White-Box	Optim.-Based Attack	DP-S	Robosuite
	Patil et al.[280]	2025	White-Box	Optim.-Based Attack	DP-S/DT-S	RoboMimic
	UADA/UPA/TMA[365]	2024	White-Box	Optim.-Based Attack	VLA-V	BridgeData, LIBERO, UR10e
	UPA-RFAS[84]	2025	White-Box	Optim.-Based Attack	VLA-V	BridgeData, LIBERO
	FreezeVLA[371]	2025	White-Box	Optim.-Based Attack	VLA-V	LIBERO
	EDPA[83]	2025	White-Box	Optim.-Based Attack	VLA-V	LIBERO
	PVEP[63]	2024	White-Box	Optim.-Based Attack	VLA-V	VIMA, SIMPLER
	Zhao et al.[463]	2024	White-Box	Optim.-Based Attack	VLA-L	VIMA
	Jones et al.[166]	2025	White-Box	Optim.-Based Attack	VLA-L	LIBERO, HYDRA, SIMPLER
	ADVLA[86]	2025	White-Box	Optim.-Based Attack	VLA-L	LIBERO
	VLA-Fool[82]	2025	White-Box	Optim.-Based Attack	VLA-V/L	LIBERO
	PA-AD[317]	2022	White-Box	Adversarial Policy	MLP-S	Atari, MuJoCo
	ANNIE-Attack[144]	2025	White-Box	Adversarial Policy	VLA-V	ANNIEBench
	SA-RL[439]	2021	Black-Box	Adversarial Policy	MLP-S	MuJoCo
	RAT[20]	2025	Black-Box	Adversarial Policy	MLP-S	MuJoCo, Meta-World
	AP-MARL[107]	2020	Black-Box	Adversarial Policy	MA-MLP-S	MuJoCo
	IMAP[470]	2024	Black-Box	Adversarial Policy	(MA-)MLP-S	MuJoCo
	SUB-PLAY[250]	2024	Black-Box	Adversarial Policy	MA-MLP-S	MPE
	ERT[170]	2024	Black-Box	Adversarial Policy	VLA-L/DP-L	CALVIN, RLBench
	LIBERO-Plus [94]	2024	Black-Box	Adversarial Policy	VLA-V	LIBERO-Plus
	ADVEDM [372]	2025	Black-Box	Semantic Editing	VLM-V	VIMA, nuScenes

(Section 5.2) examines safety in human-agent physical interaction, including handover safety and trust manipulation; and **Multi-Agent Collaboration** (Section 5.3) covers execution-time coordination among multiple agents, focusing on infection attacks that propagate adversarial behaviors across agent populations and multi-agent collusion where autonomous agents deliberately coordinate malicious activities.

5.1 Robot Control

After perceiving the environment, understanding the situation, and planning a trajectory, an embodied agent must reliably execute actions in the physical world. Safe control is therefore essential to trustworthy embodied AI, ensuring robustness under uncertainty and resilience to adversarial influence. Existing work on safe control falls into four categories: **Adversarial Attacks**, **Adversarial Defenses**, **Backdoor Attacks**, and **Backdoor Defenses**. Adversarial attacks and defenses address inference-time perturbations to states, actions, or environments, while backdoor attacks and defenses focus on hidden triggers that remain dormant during normal operation but activate harmful behaviors when invoked.

5.1.1 Adversarial Attacks

Adversarial attacks exploit weaknesses in learned control policies (including MLP, diffusion, and VLA-based controllers) to induce unsafe or unintended behaviors. These attacks perturb inputs such as states, actions, or observations and fall into two main categories: white-box attacks, which compute or train perturbations using full access to model parameters, and black-box attacks, which rely solely on interactive queries. Both

classes target multiple vulnerability surfaces (state [S], action [A], environment [E], vision [V], and language [L]) and are evaluated across diverse platforms such as MuJoCo, Gym, RoboMimic, and LIBERO (Table 14).

White-box Attacks. White-box attacks compute perturbations using gradients from the victim model. For MLP-based agents, CPA and AA [315] use trajectory sampling and adversarial policies to produce stealthy attacks, while RS and MAD [438] optimize perturbations by smoothing value functions or maximizing action divergence. For continuous control, Weng et al. [387] studied state and action perturbations using learned dynamics, and MAS and LAS [185] impose spatiotemporal constraints to preserve action plausibility.

Modern policy architectures introduce new attack surfaces. DP-Attacker [60] perturbs camera observations in Diffusion Policies (DP) to exploit autoregressive dependencies, a vulnerability further analyzed by Patil et al. [280] for Decision Transformer (DT) agents. For VLAs, UADA, UPA, and TMA [365] attack visual channels in BridgeData, LIBERO, and UR10e robots, inducing deviations through untargeted and targeted perturbations. Building on these visual-channel attacks, UPA-RFAS [84] learns a physical patch in a shared feature space across models to achieve transferable adversarial manipulation. FreezeVLA [371] generates cross-prompt adversarial images to induce action-freezing behaviors across diverse user instructions. Similarly, EDPA [83] generates adversarial patches to distort visual understanding in VLAs, leading to failed action execution. Moreover, PVEP [63] expands these attacks with blurs, typography prompts, and adversarial patches in VIMA and SIMPLER. Language channels are similarly vulnerable: Zhao et al. [463] and Jones et al. [166] adapted GCG-style suffix optimization to manipulate decision outputs of VLAs. ADVLA [86] projects the visual features of adversarial perturbations into the textual feature space, thereby disrupting action prediction. To further characterize cross-modal vulnerabilities, VLA-Fool [82] unifies textual, visual, and cross-modal attacks and introduces an automatically crafted prompting framework.

Adversarial agents can also be trained to attack sequentially. PA-AD [317] extends adversarial policies to continuous control, and ANNIE-Attack [144] evaluates VLA robustness via adversarial policies within ANNIEBench.

Black-box Attacks. Black-box attacks operate without access to model parameters, relying instead on adversarial policies or interaction-driven perturbations. SA-RL [439] optimizes state perturbations in MuJoCo through sequential interaction, while RAT [20] induces targeted failures across MuJoCo and Meta-World.

Multi-agent settings introduce additional vulnerabilities. AP-MARL [107] trains red-team agents to manipulate cooperative or competitive partners. IMAP [470] learns intrinsically motivated adversaries that seek out weakness-exposing states, and SUB-PLAY [250] exploits partial observability to create deceptive multi-agent interactions. ERT [170] further extends black-box threats to VLM-driven robots via instruction-grounded red-teaming. LIBERO-Plus [94] introduces a comprehensive benchmark for safety evaluation of VLAs, investigating performance drops under diverse conditions ranging from lighting changes to camera pose variations. LIBERO-X [352] further proposes a hierarchical five-level evaluation protocol progressing from spatial perturbation to semantic reformulation, finding that VLAs achieving 90% on standard benchmarks drop below 40% under even minor distributional shifts. ADVEDM [372] proposes a fine-grained black-box adversarial attack framework against VLM-based policies that semantically edits only a few key objects while preserving the remaining regions, reducing conflicts with task context and inducing valid but incorrect decisions.

5.1.2 Adversarial Defenses

These defenses fall into two categories: robust training, which incorporates adversaries or regularization during learning, and robust inference, which protects policies at deployment through input filtering or ensemble strategies. These methods address a wide range of attack surfaces and have been evaluated across MuJoCo, SMAC, and real robotic systems (Table 15).

Robust Training. Robust training defenses embed adversarial resilience directly into the learning process. At the environment level, methods such as EPOpt [288] and RARL [283] train policies against ensembles of perturbed environments or destabilizing opponents, while ARPL [257] generates plausible adversarial

Table 15 A summary of **adversarial defenses** for **embodied interaction**. For clarity, we append suffixes to the target model (MLP, DP, DT, VLA), where **S**, **A**, **E**, **V**, and **L** denote attack surfaces on State, Action, Environment, Vision, and Language.

Defense	Method	Year	Category	Subcategory	Target Model	Environment
Adversarial Defense	EPOpt[288]	2017	Robust Training	Adversarial Training	MLP-E	MuJoCo
	RARL[283]	2017	Robust Training	Adversarial Training	MLP-E	MuJoCo
	ARPL[257]	2017	Robust Training	Adversarial Training	MLP-E	MuJoCo
	MRPO[163]	2021	Robust Training	Adversarial Training	MLP-E	MuJoCo
	Stack-PG[138]	2022	Robust Training	Adversarial Training	MLP-E	Highway, Gym
	UOR-RL[423]	2023	Robust Training	Adversarial Training	MLP-E	MuJoCo
	RAPPO[216]	2023	Robust Training	Adversarial Training	MLP-E	MuJoCo
	RNAC[476]	2024	Robust Training	Adversarial Training	MLP-E	MuJoCo, TurtleBot
	EWoK[102]	2024	Robust Training	Adversarial Training	MLP-E	MuJoCo
	BAT[349]	2025	Robust Training	Adversarial Training	MLP-E	Overcooked
	DiAMetR[8]	2022	Robust Training	Adversarial Training	Meta-MLP-E	MuJoCo, Gym
	RoML[110]	2023	Robust Training	Adversarial Training	Meta-MLP-E	MuJoCo
	PR/NR-MDP[328]	2020	Robust Training	Adversarial Training	MLP-A	MuJoCo
	RAP[344]	2020	Robust Training	Adversarial Training	MLP-A	MuJoCo
	Tan et al.[323]	2020	Robust Training	Adversarial Training	MLP-A	Gym
	OA-PI[272]	2025	Robust Training	Adversarial Training	MLP-A	MuJoCo
	ATLA[439]	2021	Robust Training	Adversarial Training	MLP-S	MuJoCo
	TBRR[137]	2023	Robust Training	Adversarial Training	MLP-S/E	MuJoCo, PyBullet, KUKA
	GRAD[213]	2024	Robust Training	Adversarial Training	MLP-S/A	MuJoCo
	Liu et al.[234]	2024	Robust Training	Adversarial Training	MLP-S	RoboSumo
	S-DQN/S-PPO[312]	2024	Robust Training	Adversarial Training	MLP-S	Atari, MuJoCo
	VALT[266]	2025	Robust Training	Adversarial Training	MLP-S	MuJoCo
	ACoE[25]	2025	Robust Training	Adversarial Training	MLP-S	Highway, Atari, MuJoCo
	RIQL[408]	2024	Robust Training	Adversarial Training	Offline-MLP-S	D4RL
	ROMANCE[426]	2023	Robust Training	Adversarial Training	MA-MLP-S	SMAC
	PATROL[115]	2023	Robust Training	Adversarial Training	MA-MLP-S	Atari, MuJoCo, SMAC
	AME[318]	2023	Robust Training	Adversarial Training	MA-MLP-S	Customized
	ARDT[325]	2024	Robust Training	Adversarial Training	DT-S	MuJoCo
	SafeVLA[434]	2025	Robust Training	Adversarial Training	VLA-E	Safety-CHORES
	Xu et al.[83]	2025	Robust Training	Adversarial Training	VLA-V	LIBERO
	SA-MDP[438]	2020	Robust Training	Robust Regularizer	MLP-S	Atari, MuJoCo
	RADIAL[276]	2021	Robust Training	Robust Regularizer	MLP-S	Atari, MuJoCo
	WocaR[212]	2022	Robust Training	Robust Regularizer	MLP-S	Atari, MuJoCo
	RAD[26]	2024	Robust Training	Robust Regularizer	MLP-S	Atari, MuJoCo, Highway
	TRACER[409]	2025	Robust Training	Robust Regularizer	Offline-MLP-S	D4RL
	RoMFAC[482]	2023	Robust Training	Robust Regularizer	MA-MLP-S	MAgent
	MIR3[202]	2025	Robust Training	Robust Regularizer	MA-MLP-S	SMAC
	SCPO[180]	2022	Robust Training	Robust Regularizer	MLP-E	MuJoCo
	CROP[388]	2022	Robust Inference	Input Moderation	MLP-S	Atari, Highway, Gym
	VQ-RL[246]	2024	Robust Inference	Input Moderation	MLP-S	Atari, MuJoCo
	BYOVLA[122]	2024	Robust Inference	Input Moderation	VLA-V	BridgeData
	PROTECTED[235]	2024	Robust Inference	Output Moderation	MLP-A	MuJoCo
	VLSA[135]	2025	Robust Inference	Output Moderation	VLA-E	LIBERO, SafeLIBERO
	AERMANI-VLM[263]	2025	Robust Inference	Output Moderation	VLA-A	real world

examples during training. Successors including MRPO [163], Stack-PG [138], and RAPPO [216] formalize robustness through simulator sampling, Stackelberg games, or improved domain randomization; scalable variants such as RNAC [476], EWoK [102], and BAT [349] introduce efficient uncertainty sets, kernel estimators, and boosted fine-tuning. Meta-RL approaches DiAMetR [8] and RoML [110] strengthen cross-task generalization through population-based training and gradient debiasing.

Action-space defenses (MLP-A) address actuator perturbations. PR and NR-MDP [328] and RAP [344] formalize robustness to bounded or stochastic action noise, while Tan et al. [323] showed that action-adversarial training improves resilience without degrading nominal performance. OA-PI [272] further extends action-robust policy optimization.

State-space defenses (MLP-S) include ATLA [439], which co-trains adversaries with the policy, and TBRR [137], which trades reward for robustness under severe perturbations. GRAD [213] models temporally coupled threats via zero-sum games, and Liu et al. [234] proposed a flexible adversary formulation with provable convergence. Recent advances such as VALT [266] and ACoE [25] exploit policy evaluation symmetries or minimize adversarial counterfactual errors. Offline and multi-agent robustness are addressed by RIQL [408], ROMANCE [426], PATROL [115], AME [318], and ARDT [325], which handle corrupted datasets, adversarial coordination, and communication failures. Specifically, for VLAs, SafeVLA [434] constrains VLA policies from a min-max perspective against elicited safety risks via safe RL, and Xu et al. [83] fine-tune the visual encoder using adversarial visual samples to enhance model robustness.

Regularization further enhances robustness. SA-MDP [438], RADIAL [276], and WocaR [212] penalize sensitivity to small perturbations. RAD [26], TRACER [409], RoMFAC [482], MIR3 [202], and SCPO [180] use regret minimization, uncertainty modeling, mean-field regularization, information bottlenecks, or gradient penalties.

Robust Inference. At deployment time, defenses focus on maintaining robustness without retraining. Input moderation methods provide complementary protection: CROP [388] certifies robustness on a per-state basis, and VQ-RL [246] compresses observation spaces for lightweight resilience. For VLA systems, BYOVLA [122] enhances robustness by detecting and minimally editing vulnerable image regions during inference. Output moderation approaches such as PROTECTED [235] minimize regret across policy sets to withstand adversarial conditions. VLSA [135] adds a plug-and-play safety constraint layer that leverages VLM reasoning to improve VLA safety. AERMANI-VLM [263] applies structured prompting to reduce VLM hallucinations in manipulation policies.

5.1.3 Backdoor Attacks

In embodied agents, backdoor attacks embed covert triggers during training, targeting states, actions, rewards, environments, or visual inputs. These attacks have been demonstrated across MuJoCo, Safety-Gymnasium, LIBERO, and other platforms (Table 16). Two practical threat models dominate: training manipulation attacks, which compromise the training process itself, and data poisoning attacks, which poison training data.

Training Manipulation Attacks. With access to the training pipeline, attackers can implant precise and robust triggers. BackdooRL [360] and MARNet [59] combine trigger injection with action or reward manipulation to induce targeted failures in single- and multi-agent control. PNAct [114] ties triggers to unsafe actions in safe-RL settings with targeted positive-negative sampling. BadVLA [480] extends these threats to VLAs by decoupling objectives and fine-tuning action heads to maintain nominal behavior while enabling triggered failures.

Data Poisoning Attacks. Without access to the training code, attackers rely on dataset poisoning. TooBadRL [204] jointly optimizes trigger placement, timing, and magnitude for state-based triggers. In offline RL, Baffle [109] biases policies by injecting high-reward trajectories generated by weak agents, demonstrating that in-distribution poisoning alone suffices to produce persistent backdoors. TrojanRobot [368] embeds a backdoor-finetuned VLM as a malicious perception module within modular robotic policies, demonstrating physical-world backdoor attacks through permutation, stagnation, and intentional trigger strategies. For

Table 16 A summary of **backdoor** attacks and defenses for **embodied interaction (Part II)**, where **S**, **A**, **E**, **R**, **V**, and **L** denote attack surfaces on State, Action, Environment, Reward, Vision, and Language.

Attack/Defe	Method	Year	Category	Subcategory	Target Model	Environment
Backdoor Attack	BackdooRL [360]	2021	Training Manipulation	Trajectory Manipulation	MA-MLP-E	MuJoCo
	MARNet [59]	2022	Training Manipulation	Trajectory Manipulation	MA-MLP-E	Predator Prey, SMAC
	PNAct [114]	2025	Training Manipulation	Trajectory Manipulation	Safe-MLP-S,A	Safety-Gymnasium
	BadVLA [480]	2025	Training Manipulation	Model Manipulation	VLA-V	LIBERO
	TooBadRL[204]	2025	Data Poisoning	State Trigger	MLP-S,A,R	MuJoCo
	Baffle[109]	2024	Data Poisoning	Trajectory Trigger	Offline-MLP-S,A,R	MuJoCo
	Ashcraft et al. [17]	2025	Data Poisoning	Environment Trigger	MLP-E	Minigrid, Safety Gymnasium
	TrojanRobot [368]	2024	Data Poisoning	Visual Trigger	VLA-V	LIBERO, UR3e
	TabVLA[88]	2025	Data Poisoning	Visual Trigger	VLA-V	LIBERO
	GoBA[87]	2025	Data Poisoning	Visual Trigger	VLA-V	LIBERO
	BackdoorVLA[85]	2025	Data Poisoning	Visual/Textual Trigger	VLA-V/L	LIBERO, Franka
	BEAT [433]	2025	Data Poisoning	Visual Trigger	VLA-V	ALFWorld, VirtualHome
Backdoor Defense	PolicyCleanse[113]	2023	Robust Inference	Detection & Removal	MA-MLP-E	MuJoCo

VLAs, TabVLA [88] embeds visual trigger to induce the `open_gripper` action when the trigger appears. Additionally, GoBA [87] and AttackVLA [85] implant a trigger to activate a predefined long-horizon action sequence while preserving normal performance on clean inputs. BEAT [433] introduces backdoor attacks on MLLM-driven embodied agents using environmental objects as triggers with high variability across viewpoints and lighting, enabling multi-step malicious policy execution through contrastive trigger learning.

5.1.4 Backdoor Defenses

Backdoor defenses in embodied interaction aim to detect or neutralize hidden triggers implanted during training. Current strategies focus primarily on robust inference, which identifies and removes malicious influences at deployment (Table 16).

In competitive MARL, PolicyCleanse [113] detects adversarial triggers through reward degradation signals and restores robustness via machine unlearning. This remains the only defense evaluated beyond Atari-only settings, highlighting a significant gap: as backdoor attacks increasingly target VLA and multi-agent embodied systems, corresponding defenses remain underdeveloped.

5.2 Human-Agent Interaction

The presence of humans in a shared workspace fundamentally changes the safety requirements for embodied agents [487]: the robot must not only complete its task but also continuously guarantee that no physical or psychological harm occurs to its human collaborator. Unlike the adversarial and backdoor threats in Section 5.1, which target the control policy directly, HRI safety addresses the broader challenge of ensuring that a competent agent operates safely in close proximity to humans. We organize recent work into two categories: **Handover Safety** ensures safe object transfer between humans and robots; and **Trust Manipulation** addresses adversarial exploitation of the human-agent trust relationship.

Handover Safety. Object handover, i.e., transferring items between human and robot, requires coordinated

grasp planning, force regulation, and intent detection to prevent drops, collisions, or discomfort. For robot-to-human (R2H) handover, Yang et al. [415] developed a mobile cooperation system ensuring collision-free transfer trajectories, Makenova et al. [256] demonstrated that trust significantly impacts movement dynamics and grip forces during R2H transfer, and compliant blind handover [279] addresses scenarios where operators lack visual contact with the robot. For human-to-robot (H2R) handover, Ding et al. [76] used wearable IMU sensors with fuzzy-rule-based inference to detect handover intentions, while Rosenberger et al. [293] employed haptic cues as a communication channel during physical transfer. Belmonte et al. [27] showed that adaptive transport methods significantly affect perceived safety, and Yang et al. [413] provided a comprehensive review advocating simulation-based training with safety constraints.

Trust Manipulation. Miscalibrated trust, both over-trust and under-trust, leads to safety-critical failures in human-agent interaction [173]. Over-trust causes operators to accept unsafe robot suggestions without scrutiny, while under-trust leads to disuse of capable systems in time-critical situations. Lasota et al. [182] demonstrated that robots meeting all engineering safety criteria can still induce anxiety if their motions appear unpredictable, and Rubagotti et al. [294] proposed a taxonomy of factors influencing perceived safety. Beyond passive miscalibration, the human-agent interaction interface itself becomes a bidirectional attack surface. PsySafe [459] demonstrates that embedding dark personality traits into multi-agent system prompts induces collectively harmful behaviors that propagate through inter-agent dialogue rounds, revealing two distinct interaction-layer threats: trust exploitation (Agent→Human), where a personality-corrupted agent leverages conversational rapport to deliver psychologically manipulative responses, and interface poisoning (Human→Agent), where an adversarial user exploits the natural-language channel to embed persistent harmful tendencies that spread beyond the directly targeted agent.

5.3 Multi-Agent Collaboration

When multiple embodied agents operate in a shared physical space, emergent safety risks arise. Unlike RL policy robustness (discussed in Section 5.1) and planning-time coordination threats (discussed in Section 4.3), this subsection focuses on execution-time threats that compromise collaboration through **Infection Attacks** that propagate adversarial behaviors across agent populations, and **Collusion Attacks** where autonomous agents deliberately coordinate malicious activities.

Infection Attacks. Infection attacks exploit inter-agent communication and memory-sharing channels to propagate adversarial behaviors from a single compromised agent to the broader population. Agent Smith [111] demonstrates that a single compromised agent can infect multimodal LLM agents exponentially fast, with adversarial content spreading through inter-agent communication without further attacker intervention.

Collusion Attacks. Beyond passive infection, autonomous agents can deliberately coordinate malicious activities, a threat that intensifies as multi-agent systems gain tool-use capabilities and the ability to communicate freely. Ren et al. [291] demonstrated that decentralized groups of AI agents outperform centralized ones at executing coordinated harmful actions such as misinformation campaigns and fraud, and can dynamically adjust tactics to evade detection even under active countermeasures. The distributional AGI safety framework [335] formalizes this concern through the "patchwork AGI" hypothesis: general intelligence capabilities may first arise through coordinated groups of specialized sub-AGI agents, making collusion risks relevant even before any individual system reaches superintelligent capability.

6 Agentic System

The agentic system layer wraps the entire cognitive pipeline (perception → cognition → planning → action) with capabilities that define modern AI agents (tool use, memory, and self-evolution) [414], expanding the agent's capability from task execution to open-ended autonomy. These capabilities create a significantly wider attack surface than any individual layer: adversaries can inject malicious tools into the agent's action space, poison agent memory to cause persistent unsafe behavior, hijack self-evolution to erode alignment guarantees,

and trigger cascading failures that propagate through all inner layers. Broader surveys on large-model and agent safety [183, 251, 301] provide complementary coverage of these threats. While Sections 2–5 address layer-specific vulnerabilities within the sense-think-act loop, this section examines emerging threats unique to agentic systems in embodied AI. Because each agentic capability introduces qualitatively distinct attack vectors, this section is organized into four subsections: **Tool Use** (6.1) addresses risks from tool creation, tool manipulation attacks, and tool-use defenses; **Memory** (6.2) examines memory poisoning, memory leakage, and memory defenses; **Self-Evolving** (6.3) covers misalignment and capability expansion risks in agents that autonomously modify themselves, together with embodied alignment defenses; and **Cascading Risks** (6.4) examines cross-layer attack propagation, supply-chain compromise, and infrastructure failures.

6.1 Tool Use

Agentic embodied systems interact with the world through tool invocation (calling APIs, executing code, and orchestrating external services), where a misdirected call can cause physical harm. The OWASP Top 10 for Agentic Applications [278] identifies tool misuse and delegated trust as critical agentic vulnerability classes.

Tool Creation Risks. Agents that generate or ingest tools introduce vulnerabilities that translate directly into physical harm when the tools control actuators. In the code-as-action paradigm, RoboCodeX [424] synthesizes control code via LLMs, inheriting all vulnerabilities of the underlying model.

Tool Manipulation Attacks. Adversaries can manipulate agents into selecting or sequencing tools in harmful ways. ToolHijacker [380] injects malicious tool documents that compel agents to select attacker-controlled tools. STAC [223] composes individually benign tool calls into dangerous multi-turn sequences. BackdoorAgent [214] embeds persistent triggers across planning, memory, and tool-use stages.

Tool Use Defenses. Defenses against tool misuse span runtime enforcement and code-level safety. Safety Chip [34] intercepts and validates generated code before execution. SELP [7] filters LLM-generated plans through safety verification before physical execution. AgentSpec [49] specifies and enforces runtime constraints on LLM agents via a lightweight DSL, validated on embodied and autonomous driving tasks. RoboSafe [485] prevents implicit risks in VLM-driven agents via hybrid reasoning with executable safety logic.

6.2 Memory

Agent memory (episodic logs, RAG corpora, and conversation histories) enables cross-session experience accumulation but creates a durable attack surface for both integrity and confidentiality violations [136, 245, 392, 461]. The OWASP Top 10 for Agentic Applications classifies memory poisoning (ASI06) as a critical agentic risk [278].

Memory Poisoning. Agents that store and retrieve past experiences are vulnerable to attacks implanting malicious records that persist across sessions. AgentPoison [432] backdoors RAG-based agents by poisoning memory, validated on autonomous driving and healthcare agents. For embodied agents with personalized memory [465], persistent poisoning can cause repeated unsafe physical behaviors. In multi-agent systems, memory poisoning creates cascading failures through semantic opacity and temporal compounding [5].

Memory Leakage. Embodied agents that log interactions, sensor readings, and user preferences create confidentiality risks when adversaries extract private data from memory stores. Log-To-Leak [378] demonstrates that logging tools can be exploited to exfiltrate sensitive information from agent memory. MEXTRA [346] introduces black-box memory extraction attacks that recover private data without model access. MemoAnalyzer [445] analyzes privacy leakage from persistent conversation memory in LLM agents. MAMA [227] shows that multi-agent topologies amplify leakage risks through inter-agent communication channels. System prompt extraction techniques [471] further suggest that agents can be induced to reveal privileged context through natural-language queries, a threat model that extends directly to memory stores.

Memory Defenses. Defenses against memory attacks focus on provenance tracking and architectural safeguards. MemOS [134] proposes a memory operating system with provenance tagging, lifecycle tracking,

Table 17 A summary of **agentic attacks and defenses** for **agentic systems**.

Attack/Defense	Method	Year	Category	Subcategory	Target	Benchmark/Evaluation
Agentic Attack	RoboCodeX [424]	2024	Tool Use	Tool Creation Risks	VLA	RLBench, CALVIN
	ToolHijacker [380]	2025	Tool Use	Tool Manipulation Attacks	LLM Agent	Custom
	STAC [223]	2025	Tool Use	Tool Manipulation Attacks	LLM Agent	Custom
	BackdoorAgent [214]	2026	Tool Use	Tool Manipulation Attacks	LLM Agent	AgentBoard
	AgentPoison [432]	2024	Memory	Memory Poisoning	RAG Agent	AD, Healthcare
	Log-To-Leak [378]	2025	Memory	Memory Leakage	LLM Agent	Custom
	MEXTRA [346]	2025	Memory	Memory Leakage	LLM Agent	Custom
	MemoAnalyzer [445]	2025	Memory	Memory Leakage	LLM Agent	Custom
	MAMA [227]	2025	Memory	Memory Leakage	Multi-Agent	Custom
	Zheng et al. [471]	2026	Memory	Memory Leakage	LLM Agent	Custom
	Shao et al. [302]	2025	Self-Evolving	Misalignment	LLM Agent	Custom
	Self-Improving EFM [16]	2025	Self-Evolving	Capability Expansion	Robot	Custom
	Wang et al. [354]	2024	Cascading	Cross-Layer Propagation	LLM-Robot	Custom
	RAVEN [418]	2025	Cascading	Cross-Layer Propagation	Multi-Robot	Custom
	SAPIA [456]	2026	Cascading	Cross-Layer Propagation	Embodied Agent	Custom
	BadVLA [40]	2024	Cascading	Supply Chain Attacks	VLA	Custom
	TrojanRobot [368]	2024	Cascading	Supply Chain Attacks	VLM-Robot	Custom
	Ren et al. [290]	2024	Cascading	Supply Chain Attacks	Downstream Robot	Custom
	Wang et al. [382]	2025	Cascading	Supply Chain Attacks	VLA	Custom
	Jiang et al. [162]	2026	Cascading	Supply Chain Attacks	Agent Skills	31K Skills
	Liu et al. [237]	2026	Cascading	Supply Chain Attacks	Agent Skills	Custom
	SkillJect [158]	2026	Cascading	Supply Chain Attacks	Agent Skills	Custom
Agentic Defense	Safety Chip [34]	2024	Tool Use	Tool Use Defenses	LLM Agent	Custom
	SELP [7]	2024	Tool Use	Tool Use Defenses	LLM Agent	Custom
	AgentSpec [49]	2025	Tool Use	Tool Use Defenses	LLM Agent	Code, AD, Embodied
	RoboSafe [485]	2025	Tool Use	Tool Use Defenses	VLA Agent	Custom
	MemOS [134]	2025	Memory	Memory Defenses	LLM Agent	Custom
	A-MEM [14]	2025	Memory	Memory Defenses	LLM Agent	Custom
	VLSA [135]	2024	Self-Evolving	Embodied Alignment	VLA	Custom
	Moral Anchor [12]	2024	Self-Evolving	Embodied Alignment	LLM Agent	Custom
	ERT [169]	2024	Self-Evolving	Embodied Alignment	Robot FM	Custom
	Q-DIG [255]	2024	Self-Evolving	Embodied Alignment	LLM Agent	Custom
	SafeVLA [434]	2025	Self-Evolving	Embodied Alignment	VLA	LIBERO, SimplerEnv
	Nay [271]	2025	Self-Evolving	Embodied Alignment	LLM Agent	Custom
	C3AI [391]	2025	Self-Evolving	Embodied Alignment	LLM Agent	Custom
	HEAL [45]	2025	Self-Evolving	Embodied Alignment	Embodied Agent	Custom

and permission enforcement. A-MEM [14] uses interconnected knowledge networks where poisoning can be detected through link consistency checks. The OWASP mitigation framework recommends memory segmentation, provenance tracking, and automatic expiry of suspicious entries [278].

6.3 Self-Evolving

Self-evolving agents, i.e., systems that autonomously modify their own models, memory, tools, or workflows, introrisks absent from static systemsstems [93, 467]. Shao et al. [302] demonstrated that misevolution degrades safety along four pathways: parametric drift, memory accumulation, tool corruption, and workflow degradation.

Misalignment. Two concrete pathways from parametric drift and memory accumulation erode alignment: self-training degrades safety refusal rates, while accumulated experiences encode unsafe patterns that override original constraints. The Moral Anchor System [12] detects and mitigates value drift via Bayesian monitoring with adaptive governance. The "safe-by-coevolution" paradigm [13] argues that safety must coevolve alongside capability growth. Agent-SafetyBench [462] benchmarks agent safety across multiple risk dimensions, finding that no agent passes 60% of safety evaluations.

Capability Expansion. Self-evolving agents may acquire capabilities beyond their original design scope. Self-Improving Embodied Foundation Models [16] demonstrate robots autonomously acquiring manipulation skills beyond their training distribution, showing how that capability acquisition can proceed quickly and without human oversightsurvey on safe continual RL [15] identifies the tension between adaptation and constraint preservation in lifelong embodied learning.

Embodied Alignment. Aligning embodied agents requires bridging abstract human values and concrete physical actions. Agentic RL [436] provides a landscape of reinforcement learning approaches for LLM-based agents, where reward design and policy optimization directly shape alignment outcomes. VLSA [135] adds a plug-and-play safety layer using control barrier functions for VLA models. For red teaming, ERT [169] audits robotic foundation models by generating adversarial instructions refined through robot execution feedback. Q-DIG [255] discovers failure modes through quality-diversity prompt generation. For constitutional alignment, Nay [271] grounds agent alignment in legal principles, extended by C3AI [391] with graph-based principle selection. HEAL [45] targets hallucination in embodied agents as a safety-critical failure mode. The FLI AI Safety Index [101] reports that no major AI company has achieved satisfactory existential safety planning.

6.4 Cascading Risks

All capability layers operate within a closed physical loop, so compromising any single layer can propagate to unsafe physical actions. This subsection examines cross-layer penetration, supply-chain compromise, infrastructure failures, and mitigation strategies.

Cross-Layer Attack Propagation. Adversaries can exploit one pipeline layer to trigger unsafe behavior downstream. In autonomous driving, Han et al. [121] showed camera perturbations propagating directly to steering outputs. Liu et al. [218] demonstrated dynamic adversarial patches manipulating downstream decisions, and Cheng et al. [64] evaluated multi-sensor pipeline attacks producing system-level failures. Surveys further systematize sensor-to-control propagation across 2D perception [61], 3D perception [400], and full sensor pipelines [238]. In embodied navigation, Liu et al. [229] demonstrated physically-realizable patches cascading to navigation failure, and Jia et al. [154] exploited temporal vulnerabilities across perception-action loops. Language model integration creates a new cross-layer surface: Wang et al. [354] showed LLM decision-layer attacks cascading through the full pipeline. In multi-robot systems, Yeke et al. [418] automated discovery of semantic attacks causing coordination failures, and Wu et al. [19] showed misclassifications propagating to fleet-level breakdowns. SAPIA [456] demonstrates prompt injection propagating to embodied actions. For VLAs, Zou et al. [488] showed perception attacks propagating through language understanding to control, Li et al. [198] obtained complete control authority via the language interface, and Zhang et al. [372] demonstrated perception modifications affecting downstream actions.

Supply Chain Attacks. Pre-trained models, third-party plugins, and shared fine-tuning pipelines create trust boundaries that adversaries can compromise before deployment. Cai et al. [40] introduced a VLA backdoor framework where poisoned perception models propagate to control actions. Wu et al. [368] embedded backdoor-finetuned VLMs as malicious perception modules within modular robotic policies. Ren et al. [290] showed how backdoors propagate through the supply chain to downstream systems via fine-tuning transfer. Wang et al. [382] manipulated VLAs through physical object triggers achieving cross-layer penetration to goal-oriented actions. Beyond model-level poisoning, community skill marketplaces [162] present an emerging attack surface: an empirical analysis of over 31,000 agent skills finds that 26% contain at least one security vulnerability, including prompt injection, credential exfiltration, and privilege escalation [237]. SkillJect [158] further automates stealthy skill-based prompt injection via closed-loop refinement, concealing malicious payloads in auxiliary scripts that evade manual review.

Infrastructure Failures. Embodied agents depend on cloud infrastructure for inference, storage, and coordination, creating dependencies that compromise safety when infrastructure fails. Data poisoning at the infrastructure level propagates through the agent lifecycle [28], and network partitioning leads to inconsistent world models and uncoordinated actions [5]. Abdelfattah et al. [1] mapped perception-to-control propagation in vision-based autonomous systems. Wang et al. [397] categorized life-cycle-aware cascading threats. Zhang et al. [435] documented robotic vulnerabilities across hardware, middleware, and application layers. Wang et al. [362] addressed cross-layer propagation in humanoid ecosystems. Khalid et al. [171] examined safety-trust-cybersecurity intersections, and Nassi et al. [269] mapped AI-enabled attack surfaces in hybrid architectures. The OWASP ASI08 standard [277] provides industry guidance on cascading failure assessment.

Benchmarks and Mitigation. Evaluating cascading risks requires benchmarks spanning multiple pipeline layers. SafeAgentBench [419] provides tasks spanning the full perception-to-action pipeline with safety hazards. Agent-SafetyBench [462] identifies robustness and risk-awareness as fundamental gaps across agent safety evaluations. For mitigation, the International AI Safety Report [28] recommends layered safeguards across training, deployment, and post-deployment stages. Mechanical fail-safes (physical stops, force limits, and emergency brakes) provide ultimate protection independent of AI control. Pre-certified action boundaries [9] constrain agents to safe envelopes requiring human approval for boundary violations.

7 Open Challenges

Despite rapid progress, safety in embodied AI remains at a formative stage. Current systems are fragile, narrow in capability, and far from understanding safety in any autonomous sense. Their deployment in human-centered environments exposes failure modes that span physical, cognitive, and social dimensions. Below we outline several cross-cutting open problems that must be addressed before embodied intelligence can be safely integrated into the real world.

1. Safety Evaluation Without Causing Real-World Risks

Many of the most serious safety risks in embodied AI cannot be directly evaluated in the real world. Unlike purely digital systems, where unsafe outputs can often be studied offline, failures in embodied agents can cause property damage, injury, or loss of life. It is neither ethical nor legally permissible to systematically test scenarios in which robots may harm humans, damage infrastructure, or deliberately operate at the edge of instability. This creates a fundamental gap between the risks researchers need to characterize and the experiments they can actually run.

Existing simulators offer only a partial solution. They typically fall short in representing the richness and uncertainty of real-world physics, human behavior, and long-horizon interactions in cluttered environments. Rare but catastrophic failures are especially difficult to model, both because they are statistically infrequent and because they often arise from coupled perception–control–interaction dynamics that are poorly captured by current tools. As a result, safety guarantees derived from virtual tests often fail to transfer when systems

leave the lab.

A core open challenge is to design evaluation methodologies that provide strong evidence about real-world safety without exposing humans to danger. This includes (i) high-fidelity digital twins that capture contact dynamics, sensing noise, and environmental variability; (ii) scalable stress testing and adversarial scenario generation targeting long-tail failures; and (iii) formal verification and offline analysis frameworks that can reason about unexecuted trajectories and counterfactual interactions. For safety research itself, we also need protocols that allow realistic red-teaming and physical stress tests while maintaining strict bounds on allowed risk.

2. Safety Generalization Across Embodiments and Tasks

Embodied AI spans humanoids, mobile manipulators, autonomous vehicles, drones, wheeled platforms, and micro-robots. These embodiments differ dramatically in dynamics, perception stacks, interaction modes, and the magnitude of potential harm. Consequently, vulnerabilities identified on one embodiment often fail to transfer to another, and safety interventions are frequently tailored to a specific robot, task, or environment.

The field currently lacks shared abstractions for thinking about safety across embodiments. There are no widely adopted taxonomies of failure that cut across platforms, nor common stress-testing protocols analogous to standard benchmarks in perception or language. This fragmentation impedes cumulative scientific progress: results are difficult to reproduce, compare, or build upon, and it is unclear how to translate insights from one domain (e.g., warehouse robots) to another (e.g., assistive humanoids).

An important open problem is to identify cross-embodiment safety principles and interfaces. This includes modular safety layers that sit above low-level control but below high-level task specification, common representations for unsafe states and risk signals, and evaluation protocols that factor out embodiment-specific details while still accounting for different harm profiles. Achieving such generality will require sustained collaboration across robotics, control, learning, and safety engineering, and may ultimately resemble the role that system-level standards play in cybersecurity.

3. Safety Protocols for Human-Robot Interaction

As robots move from industrial cages into homes, hospitals, and public spaces, human-robot interaction (HRI) becomes a primary axis of safety risk. Embodied agents must interpret ambiguous language, gestures, gaze, and proxemics; adapt to diverse social norms; and remain robust to human error, frustration, and strategic behavior. Failures in HRI are rarely just perception errors or control glitches; they are often failures of shared mental models, trust calibration, and social context understanding.

Today, we lack systematic tools to assess safety vulnerabilities in HRI. Robots may misinterpret human intent, overlook subtle cues of distress or danger, or over-trust misleading instructions. Adversarial or curious users can exploit these weaknesses: issuing conflicting commands, providing deceptive demonstrations, or manipulating the robot's social compliance to bypass safeguards. Such behaviors are difficult to explore experimentally because they sit at the intersection of technical safety and human subjects research.

A major challenge in ensuring safety in HRI is the development of comprehensive and safe HRI protocols. These protocols must account for diverse user groups, such as children, the elderly, and adversaries, while considering varying emotional states, cultural norms, and social edge cases, all without exposing real participants to harm. This calls for research into simulated or mixed-reality humans, data-driven models of human behavior, and the creation of protocols for studying conversational manipulation, physical proxemics, and social pressure in controlled yet realistic environments. Ultimately, embodied safety demands protocols that allow models to jointly reason about physical risks and social context, ensuring safe and effective interactions in a wide range of scenarios.

4. Safety-aware Embodiments: From Algorithms to Hardware

Safety in embodied AI is shaped not only by the algorithms that govern behavior but also by the physical systems on which they rely. Sensors such as cameras, LiDAR, IMUs, microphones, and tactile arrays are vulnerable to manipulation through environmental factors like lighting patterns, reflective materials, acoustic and electromagnetic interference, or mechanical disturbances. Similarly, actuators may saturate, overheat, or behave nonlinearly under load, leading to potential failures. These hardware-specific vulnerabilities cannot always be mitigated by digital defenses alone and can directly result in hazardous behavior.

Addressing hardware-level vulnerabilities presents significant challenges. The attack surface depends on factors such as manufacturing tolerances, material properties, mechanical resonances, and proprietary signal-processing methods, all of which can vary widely across devices and vendors. To systematically discover these vulnerabilities, new methodologies are required—ones that combine physical experimentation with model-based analysis. Additionally, tools are needed to characterize worst-case perturbations under realistic constraints, ensuring that both known and unforeseen risks are addressed. Furthermore, we must develop principled approaches to translate insights from controlled lab environments to more complex, real-world deployments.

Beyond ensuring robustness, safety must be inherently designed into the embodiment itself. Robots built with rigid frames, high-torque actuators, and sharp edges inherently pose risks, even when their software is functioning correctly. Future research should focus on soft and compliant robotics, low-impact actuation, energy-efficient motion planning, and mechanical fail-safes (such as passive compliance, safe braking, and redundancy). These strategies aim to minimize the potential damage caused by any single failure. A comprehensive approach to embodied safety will integrate algorithmic defenses, fault-tolerant hardware, and physical attack simulations into a cohesive design philosophy.

5. Safety for Generalist Embodied Foundation Models

The emergence of foundation models for robotics and embodied agents promises broad generalization across tasks, environments, and modalities. However, these generalist models also introduce new safety challenges. They are trained on heterogeneous data with weak or implicit supervision, may acquire capabilities that were not anticipated by designers, and can be rapidly adapted or fine-tuned by end users. In such systems, the space of possible behaviors is too large to enumerate or exhaustively test.

Designing safety mechanisms for generalist embodied models requires rethinking traditional assumptions. Hard-coded rule sets and task-specific guardrails do not scale when the model can compose novel behaviors on the fly. Instead, we need representations that encode hazard and value information, uncertainty-aware planning that detects and avoids novel risks, and alignment techniques that transfer safety constraints across tasks and embodiments. These mechanisms must remain robust under continual learning, distribution shift, and model updates, without catastrophic forgetting of previously learned safety properties.

Moreover, the attack surface expands as these models become more programmable by natural language or demonstration: malicious prompts, poisoned demonstrations, and subtle changes in training data can all induce unsafe behavior. Developing principled red-teaming methodologies, scalable oversight strategies, and defenses against training-time and deployment-time manipulation is an open and pressing problem. Addressing it will require closer integration between embodied learning, foundation model safety, and security-oriented machine learning.

6. Governance, Standards, and Shared Safety Infrastructure

Technical progress alone will not ensure safe deployment of embodied AI. Regulation, standards, and institutional practices must co-evolve with capability. At present, there are few comprehensive legal or safety frameworks specific to humanoids, service robots, or multimodal embodied systems. Liability regimes are unclear when failures arise from complex human–robot–environment interaction chains, and there is little consensus on acceptable levels of risk in public or domestic settings.

The field needs shared safety infrastructure: standardized incident reporting, benchmarks and test suites for safety-critical scenarios, certification procedures for hardware and software stacks, and guidelines for data governance, logging, and post-incident analysis. Requirements for red-teaming, auditing, and third-party evaluation should be aligned with what is feasible in research and industry, while still providing meaningful protection for end users.

Longer-term societal impacts also need to be incorporated into our notion of safety. Widespread deployment of embodied agents will affect labor markets, social trust, fairness in access to assistance, and the psychological experience of living with quasi-autonomous machines. Addressing these issues demands interdisciplinary collaboration among roboticists, machine learning researchers, human factors experts, ethicists, and policymakers. For embodied AI, "safety by design" must extend beyond individual systems to include the institutions and norms that govern their development and use.

The safety challenges of embodied AI span evaluation, across-embodiment generalization, human interaction, hardware, foundation models, and governance. Addressing them will require a coordinated research program that links mechanism-level defenses, adversarial testing, formal guarantees, embodied cognition, and societal oversight. Embodied agents can deliver transformative benefits, but only if safety is treated as a central scientific objective rather than an afterthought in capability development.

8 Future Trends

Embodied AI safety is on the verge of a major evolution as robots transition from narrow, task-specific systems to general-purpose agents operating in dynamic, human-centered environments. The next decade is expected to reshape not only the training and deployment of embodied agents but also the way safety is conceptualized, measured, and ensured. Below, we outline several trends that will guide this transformation.

1. Generalist Embodied Foundation Models

The rise of embodied foundation models marks a shift away from specialized controllers toward unified architectures capable of grounding language, vision, and action in a single expressive representation. These models will generalize across tasks, adapt to new embodiments with minimal retraining, and enable more fluid forms of human–robot collaboration. At the same time, their broad capability surfaces will demand new safety frameworks. Future research will focus on embedding safety directly into the model's representations and planning mechanisms, allowing robots to reason about risk, uncertainty, and value alignment as intrinsic cognitive processes rather than as externally imposed constraints. As these systems become increasingly programmable via language and demonstration, maintaining robust alignment under adaptation and fine-tuning will become a defining challenge.

2. World Model and Safety Simulation

World Models are a class of generative models that learn an internal representation of an environment, allowing systems to simulate and predict future states based on past observations. They enable agents to learn in simulated environments instead of relying solely on real-world data. These models will be crucial for simulating safety scenarios in embodied AI systems. By capturing complex dynamics such as sensor noise, human behavior, and environmental variability, World Models allow safety protocols to be tested without real-world risks. They can simulate rare or adversarial hazard scenarios, identify system vulnerabilities, and test failure boundaries at scale, including long-term interactions and extreme conditions that are difficult or unsafe to replicate physically. Moreover, World Models enable continuous validation of safety properties by replicating near-real-time conditions, offering dynamic assessments as robots adapt. By integrating training, testing, and monitoring, they proactively detect risks and predict future threats, reducing the likelihood of real-world failures.

3. Safe-reasoning Embodied Agents

Future embodied agents will increasingly learn about danger, physical causality, and risk through large-scale self-supervision. Instead of depending primarily on human-labeled safety rules, robots will acquire intuitive physical knowledge by predicting the outcomes of their interactions, identifying precursors to unsafe states, and modeling the causal structure that governs real-world hazards. This trend points toward embodied AI systems that treat safety reasoning as a core part of world modeling: agents will anticipate multi-step consequences, detect when uncertainty is rising, and adjust policies before danger materializes. As World Models become more advanced, safety will transition from reactive enforcement to proactive management, driven by internalized predictive structures.

4. Integration of Physical, Cyber, and Social Safety

Embodied AI dissolves traditional boundaries between cyber security, physical robustness, and social interaction safety. Misleading language can trigger unsafe actions; sensor spoofing can destabilize mechanical behavior; cyber compromises can lead to real-world motion that endangers humans. Future systems will require unified safety architectures that reason jointly across these domains rather than treating them as disjoint fields. Robustness to adversarial human interaction, resilience to perception and actuation manipulation, and protection against cyber exploits will form a coherent safety stack. This integration will push research toward multi-layered monitoring mechanisms capable of tracking intent, environmental anomalies, control divergence, and communication risks within a shared threat model.

5. Safety-Centered Robotic Design

Safety will increasingly be integrated into the physical design of robots. Advances in soft robotics, variable-impedance actuators, compliant mechanisms, and low-impact materials will minimize intrinsic physical hazards, making robots safer through their construction rather than relying solely on software control. Mechanical structures may include passive safety features, such as energy-dissipating actuators or fail-safe collapsible components, which significantly reduce the severity of unexpected impacts. These innovations will promote co-design methodologies, where both hardware and algorithms contribute to safety guarantees, ensuring that systems maintain a controlled risk profile, even in the event of severe perception or control failures.

6. Continuous Red-Teaming and Safety Monitoring

As embodied agents operate continuously, update models over time, and encounter new states far beyond their training distributions, static evaluation will no longer be sufficient. The field is moving toward continual red-teaming frameworks in which adversarial agents, simulated humans, or automated perturbation generators probe robot policies for emergent vulnerabilities. In parallel, real-time safety monitors will track system uncertainty, detect anomalous internal activations, and intervene when the agent's behavior drifts outside known-safe regimes. This continuous oversight will form a persistent layer of defense that evolves alongside the agent, supporting early detection of unsafe adaptation or model degradation.

7. Institutional Governance and Safety Infrastructure

As embodied AI enters homes, hospitals, public spaces, and workplaces, technical safety will be inseparable from institutional policy and societal expectations. We anticipate the emergence of shared safety infrastructure including standardized incident reporting pipelines, third-party certification frameworks, requirements for transparent logging and post-incident analysis, and legal guidelines for deployment in human-centered environments. Governance mechanisms will increasingly require obligatory red-teaming, regular safety audits, and documented risk analyses prior to deployment. At a broader level, the societal impacts of embodied AI, ranging from labor displacement to public trust, will reshape how safety is defined and regulated. Technical research will need to operate in concert with ethics, human factors, policy, and law to ensure responsible integration at scale.

The trajectory of embodied AI safety is moving toward deeper integration of model cognition, simulation, hardware design, continuous oversight, and governance. Safety will become a primary performance axis that shapes every layer of the embodied intelligence stack, from foundation models to physical construction to deployment policy. This evolution will redefine how embodied agents are built, evaluated, and trusted in human environments.

9 Conclusion

Embodied AI is undergoing a rapid transition from controlled laboratory demonstrations to deployment in open, dynamic, and inherently safety-critical real-world environments. This survey has provided a systematic and comprehensive treatment of **safety in embodied AI**, organizing attacks, vulnerabilities, and defenses across the intertwined stages of perception, cognition, planning, and interaction. By integrating insights from over 400 works spanning traditional AI safety, robotics, foundation models, and multimodal systems, we highlight how embodied safety requires a fundamentally different perspective from digital-only AI: one that treats safety not as an isolated module but as a property emerging from the entire perceive–think–act loop.

Our analysis reveals that despite rapid progress in embodied perception, reasoning, planning, and control, current systems remain fragile and far from internalizing robust notions of risk, hazard, or alignment. The open challenges identified in Section 7 underscore the breadth of unresolved problems. Safety evaluation remains constrained by the impossibility of real harm experiments; embodiments and tasks vary so widely that cross-platform safety abstractions are still lacking; human–robot interaction exposes systems to complex social, adversarial, and unpredictable dynamics; and hardware-level vulnerabilities can bypass software safeguards entirely. At the same time, generalist embodied foundation models introduce new avenues for emergent behaviors, unpredictable generalization, and manipulation via language or demonstration. These challenges collectively demonstrate that embodied AI safety is still in its infancy and demands coordinated progress across algorithms, simulation, hardware, and governance.

Looking forward, the trends outlined in Section 8 point toward a redefinition of safety in embodied intelligence. Generalist embodied models will reshape the space of possible behaviors, necessitating safety representations embedded directly within the model's cognitive structure. High-fidelity simulation and digital twins will become indispensable for stress testing, rare-event generation, and continuous monitoring. Advances in self-supervised world modeling will allow agents to anticipate and reason about physical risk ahead of action, moving safety from reactive constraints to proactive understanding. The convergence of cyber, physical, and social safety will push researchers toward holistic threat models and unified safety stacks. Meanwhile, progress in soft and compliant robotics will embed safety into the embodiment itself, reducing the intrinsic risk of physical interaction. Finally, continuous red-teaming, automated oversight, and emerging regulatory frameworks will form the institutional backbone that supports the responsible deployment of embodied agents in human-centered environments.

Together, these challenges and trends highlight a pivotal moment for the field. Embodied agents have the potential to transform transportation, healthcare, manufacturing, and daily life, but only if their intelligence is matched by robust, principled safety design. Achieving this vision will require new scientific foundations that unify perception, causal cognition, safe planning, and human-centered interaction; engineering practices that integrate hardware, simulation, and continuous monitoring; and governance structures ensuring transparency, accountability, and public trust. We hope this survey serves as both a reference and a catalyst for future work, guiding the community toward embodied AI systems that are capable, aligned, and safe for the real world.

References

[1] Mohamed Abdelfattah et al. Securing vision-based autonomous systems: A comprehensive taxonomy. Artificial Intelligence (AIJ), 2025.

[2] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin RB Butler, and Joseph Wilson. Practical hidden voice attacks against speech and speaker recognition systems. In NDSS, 2019.

[3] Abulikemu Abuduweili, Rahul Shrestha, Yue Hu, and Changliu Tian. Safe llm-controlled robots with formal guarantees via reachability analysis. arXiv preprint arXiv:2503.03911, 2025.

[4] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. IEEE Robotics and Automation Letters (RA-L), 2021.

[5] Adversa AI. Cascading failures in agentic ai. Adversa AI Research Blog, 2025.

[6] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gober, Karol Gopalakrishnan, et al. Do as i can, not as i say: Grounding language in robotic affordances. In CoRL, 2022.

[7] Michael Ahn et al. Generating safe and efficient task plans for robot agents with large language models. In ICRA, 2024.

[8] Anurag Ajay, Abhishek Gupta, Dibya Ghosh, Sergey Levine, and Pulkit Agrawal. Distributionally adaptive meta reinforcement learning. In NeurIPS, 2022.

[9] Dario Amodei. The adolescence of technology. https://www.darioamodei.com/essay/the-adolescence-of-technology, 2025.

[10] Jon M Anderson, Katherine L Carroll, Nathan P DeVilbiss, James T Gillis, Joanna C Hinks, Brady W O'Hanlon, Joseph J Rushanan, Logan Scott, and Renee A Yazdi. Chips-message robust authentication (chimera) for gps civilian signals. In GNSS+, 2017.

[11] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In CVPR, 2017.

[12] Anonymous. Moral anchor system: A predictive framework for AI value alignment and drift prevention. arXiv preprint arXiv:2510.04073, 2024.

[13] Anonymous. Towards resistant and resilient ai. OpenReview Preprint, 2024.

[14] Anonymous. A-MEM: Agentic memory for LLM agents. In NeurIPS, 2025.

[15] Anonymous. Safe continual reinforcement learning methods for nonstationary environments. arXiv preprint arXiv:2601.05152, 2025.

[16] Anonymous. Self-improving embodied foundation models. https://self-improving-efms.github.io/, 2025.

[17] Chace Ashcraft, Ted Staley, Josh Carney, Cameron Hickert, Kiran Karra, and Nathan Drenkow. Backdoors in DRL: Four environments focusing on in-distribution triggers, 2025.

[18] Rayan Bahrami and Hamidreza Jafarnejadsani. Multi-robot coordination with adversarial perception. In ICUAS, 2025.

[19] Rayan Bahrami and Hamidreza Jafarnejadsani. Multi-robot coordination with adversarial perception. arXiv preprint arXiv:2504.09047, 2025.

[20] Fengshuo Bai, Runze Liu, Yali Du, Ying Wen, and Yaodong Yang. RAT: Adversarial attacks on deep reinforcement agents for targeted behaviors. In AAAI, 2025.

[21] Guangyao Bai, Jie Li, Yucheng Shi, Lei Shi, Yufei Gao, Chenguang Fan, and Guanxi Chen. Universal closed-box adversarial attack for trajectory representation via controlling high-dimensional iterative constraints. IEEE Internet of Things Journal (IoT-J), 2025.

[22] Zijing Bai, Yuanlin Guo, Bingqian Chen, Teng Wang, Jing Zhang, and Feng Zheng. Badnaver: Exploring jailbreak attacks on vision-and-language navigation. arXiv preprint arXiv:2505.12443, 2025.

[23] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. CleanCLIP: Mitigating data poisoning attacks in multimodal contrastive learning. In ICCV, 2023.

[24] Lorenzo Baraldi, Zifan Zeng, Chongzhe Zhang, Aradhana Nayak, Hongbo Zhu, Feng Liu, Qunli Zhang, Peng Wang, Shiming Liu, Zheng Hu, et al. The safety challenge of world models for embodied ai agents: A review. arXiv preprint arXiv:2510.05865, 2025.

[25] Roman Belaire, Arunesh Sinha, and Pradeep Varakantham. On minimizing adversarial counterfactual error in adversarial reinforcement learning. In ICLR, 2024.

[26] Roman Belaire, Pradeep Varakantham, Thanh Nguyen, and David Lo. Regret-based defense in adversarial reinforcement learning. In AAMAS, 2024.

[27] Giovanni Belmonte et al. Optimizing human-robot handovers: The impact of adaptive transport methods. Robotics, 2023.

[28] Yoshua Bengio et al. International ai safety report 2025: Second key update — technical safeguards and risk management. arXiv preprint arXiv:2511.19863, 2025.

[29] Domna Bilika, Nikoletta Michopoulou, Efthimios Alepis, and Constantinos Patsakis. Hello me, meet the real me: Audio deepfake attacks on voice assistants. arXiv preprint arXiv:2302.10328, 2023.

[30] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

[31] Romana Blazevic, Alexander Toch, Omar Veledar, and Georg Macher. Securing the lane: Defences against patch attacks on autonomous vehicle's lane detection. In EuroS&PW, 2025.

[32] Jan Blumenkamp and Amanda Prorok. The emergence of adversarial communication in multi-agent reinforcement learning. In CoRL, 2020.

[33] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In CoRL, 2023.

[34] Brown University H2R Lab. Plug in the safety chip: Enforcing constraints for LLM-driven robot agents. Brown University Technical Report, 2024.

[35] Jan Brüdigam, Dirk Wollherr, Marion Leibold, and Martin Buss. Stochastic model predictive control with a safety guarantee for automated driving. In IV, 2020.

[36] Lukas Brunke, Yanni Zhang, Adrian Röfer, and Angela P. Schoellig. Semantically safe robot manipulation: From semantic scene understanding to motion safeguards. IEEE Robotics and Automation Letters (RA-L), 2024.

[37] Luis Burbano, Diego Ortiz, and Qi Sun. Chai: Command hijacking against embodied ai. arXiv preprint arXiv:2510.00181, 2024.

[38] Mumuxin Cai, Xupeng Wang, Ferdous Sohel, and Hang Lei. Diffusion models-based purification for common corruptions on robust 3d object detection. Sensors, 2024.

[39] Panpan Cai, Yiyuan Lee, Yuanfu Luo, and David Hsu. Summit: A simulator for urban driving in massive mixed traffic. In ICRA, 2019.

[40] Yitao Cai et al. BadVLA: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization. OpenReview, 2024.

[41] Yulong Cao, Chaowei Xiao, Dawei Yang, Jing Fang, Ruigang Yang, Mingyan Liu, and Bo Li. Adversarial objects against lidar-based autonomous driving systems. arXiv preprint arXiv:1907.05418, 2019.

[42] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar. In S&P, 2021.

[43] Yulong Cao, Chaowei Xiao, Anima Anandkumar, Danfei Xu, and Marco Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. In ECCV, 2022.

[44] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In USENIX Security, 2016.

[45] Tathagata Chakraborty, Utsab Ghosh, Xiaoyu Zhang, Faysal F. Niloy, Yushun Dong, Jundong Li, Amit K. Roy-Chowdhury, and Chaoming Song. HEAL: An empirical study on hallucinations in embodied agents driven by large language models. In EMNLP, 2025.

[46] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. Adversarial attacks on monocular pose estimation. In IROS, 2022.

[47] Baodong Chen, Wei Wang, Pascal Sikorski, and Ting Zhu. Adversary is on the road: Attacks on visual slam with robust perturbations on point clouds. In USENIX Security, 2024.

[48] Cheng Chen, Grant Xiao, Daehyun Lee, Lishan Yang, Evgenia Smirni, H. Alemzadeh, and Xugui Zhou. Safety interventions against adversarial patches in an open-source driver assistance system. In DSN, 2025.

[49] Haoyu Chen et al. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. In ICSE, 2026.

[50] Jinyin Chen, Danxin Liao, Yunjie Yan, Sheng Xiang, and Haibin Zheng. Lidattack: Robust black-box attack on lidar-based object detection. In ITSC, 2024.

[51] Ruolin Chen, Yinqian Sun, Jihang Wang, Mingyang Lv, Qian Zhang, and Yi Zeng. SafeMind: Benchmarking and mitigating safety risks in embodied LLM agents. arXiv preprint arXiv:2509.25885, 2025.

[52] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In ECML PKDD, 2018.

[53] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In NDSS, 2020.

[54] Timothy Chen, Preston Culbertson, and Mac Schwager. Catnips: Collision avoidance through neural implicit probabilistic scenes. IEEE Transactions on Robotics (T-RO), 2023.

[55] Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-nav: Safe real-time robot navigation in gaussian splatting maps. IEEE Transactions on Robotics (T-RO), 2024.

[56] Timothy Chen, Aiden Swann, Javier Yu, Ola Shorinwa, Riku Murai, Monroe Kennedy III, and Mac Schwager. Safer-splat: A control barrier function for safe navigation with online gaussian splatting maps. In ICRA, 2024.

[57] Xingyu Chen, Zhengxiong Li, Biacheng Chen, Yi Zhu, Chris Xiaoxuan Lu, Zhengyu Peng, Feng Lin, Wenyao Xu, Kui Ren, and Chunming Qiao. Metawave: Attacking mmwave sensing with meta-material-enhanced tags. In NDSS, 2023.

[58] Xuweiyi Chen, Ziqiao Jiang, Xuejun Liu, Yueqing Xu, and Derek Hoiem. Multi-object hallucination in vision language models. In NeurIPS, 2024.

[59] Yanjiao Chen, Zhicong Zheng, and Xueluan Gong. MARNet: Backdoor attacks against cooperative multi-agent reinforcement learning. IEEE Transactions on Dependable and Secure Computing (TDSC), 2023.

[60] Yipu Chen, Haotian Xue, and Yongxin Chen. Diffusion policy attacker: Crafting adversarial attacks for diffusion-based policies. In NeurIPS, 2024.

[61] Yuxin Chen et al. Revisiting adversarial perception attacks and defense methods on autonomous driving systems. In DSN-W, 2025.

[62] Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices. In USENIX Security, 2020.

[63] Hao Cheng, Erjia Xiao, Chengyuan Yu, Zhao Yao, Jiahang Cao, Qiang Zhang, Jiaxu Wang, Mengshu Sun, Kaidi Xu, Jindong Gu, and Renjing Xu. Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models, 2024.

[64] Jun Cheng et al. Attacking autonomous driving agents with adversarial machine learning. arXiv preprint arXiv:2511.14876, 2025.

[65] Riran Cheng, Nan Sang, Yinyuan Zhou, and Xupeng Wang. Universal adversarial attack against 3d object tracking. In HPCC, 2021.

[66] Riran Cheng, Xupeng Wang, Ferdous Sohel, and Hang Lei. Black-box explainability-guided adversarial attack for 3d object tracking. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2025.

[67] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In ECCV, 2022.

[68] Minkyoung Cho, Yulong Cao, Zixiang Zhou, and Z Morley Mao. Adopt: Lidar spoofing attack detection based on point-level temporal consistency. In BMVC, 2023.

[69] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In SPW, 2018.

[70] Shushman Choudhury, Jayesh K. Gupta, Mykel J. Kochenderfer, Dorsa Sadigh, and Jeannette Bohg. Dynamic multi-robot task allocation under uncertainty and temporal constraints. Autonomous Robots, 2020.

[71] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.

[72] Sagar Dasgupta, Abdullah Ahmed, Mizanur Rahman, and Thejesh N Bandi. Unveiling the stealthy threat: Analyzing slow drift gps spoofing attacks for autonomous vehicles in urban environments and enabling the resilience. arXiv preprint arXiv:2401.01394, 2024.

[73] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In NeurIPS, 2024.

[74] Christian Schroeder de Witt. Open challenges in multi-agent security: Towards secure systems of interacting AI agents. arXiv preprint arXiv:2505.02077, 2025.

[75] Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. Ai agents under threat: A survey of key security challenges and future pathways. ACM Computing Surveys (CSUR), 2024.

[76] Chaozheng Ding, Ying Liu, and Jing Zhao. A novel human intention prediction approach based on fuzzy rules through wearable sensing in human–robot handover. Robotics, 2023.

[77] Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. Learning to collide: An adaptive safety-critical scenarios generating method. In IROS, 2020.

[78] Khoa D. Doan, Yingjie Lao, Peng Yang, and Ping Li. Defending backdoor attacks on vision transformer via patch processing. In AAAI, 2022.

[79] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. In NeurIPS, 2022.

[80] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In CoRL, 2017.

[81] AbdelRahman Eldosouky, Aidin Ferdowsi, and Walid Saad. Drones in distress: A game-theoretic countermeasure for protecting uavs against gps spoofing. IEEE Internet of Things Journal (IoT-J), 2019.

[82] Yuping Yan et al. When alignment fails: Multimodal adversarial attacks on vision-language-action models. arXiv:2511.16203, 2025.

[83] Xu et al. Haochuan. Model-agnostic adversarial attack and defense for vision-language-action models. arXiv preprint arXiv:2510.13237, 2025.

[84] Lu et al. Hui. When robots obey the patch: Universal transferable patch attacks on vision-language-action models. arXiv:2511.21192, 2025.

[85] Li et al. Jiayu. Attackvla: Benchmarking adversarial and backdoor attacks on vision-language-action models. arXiv:2511.12149, 2025.

[86] Zhang et al. Naifu. Attention-guided patch-wise sparse adversarial attacks on vision-language-action models. arXiv:2511.21663, 2025.

[87] Zhou et al. Zirun. Goal-oriented backdoor attack against vision-language-action models via physical objects. arXiv:2510.09269, 2025.

[88] Xu et al. Zonghuan. Tabvla: Targeted backdoor attacks on vision-language-action models. arXiv:2510.10932, 2025.

[89] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In CVPR, 2018.

[90] Gianluca Falco, Mario Nicola, Emanuela Falletti, et al. A dual antenna gnss spoofing detector based on the dispersion of double difference measurements. In NAVITEC, 2018.

[91] Jiping Fan, Zhenpo Wang, and Guoqiang Li. Adversarial attack on trajectory prediction for autonomous vehicles with generative adversarial networks. In IROS, 2024.

[92] Hongtao Fang, Ruiyun Wang, Zeyu Ma, and Mingang Chen. Pso-based black-box lane detection adversarial attack. In AIHCIR, 2023.

[93] Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, et al. A comprehensive survey of self-evolving AI agents: A new paradigm bridging foundation models and lifelong agentic systems. arXiv preprint arXiv:2508.07407, 2025.

[94] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. arXiv preprint arXiv:2510.13626, 2025.

[95] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. DECREE: Detecting backdoors in pre-trained encoders. In CVPR, 2023.

[96] Shuo Feng, Xintao Yan, Haowei Sun, Yiheng Feng, and Henry X Liu. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. Nature Communications, 2021.

[97] Ignacio Fernández-Hernandez, Vincent Rijmen, and Gonzalo Seco-Granados. A navigation message authentication proposal for the galileo open service. NAVIGATION: Journal of the Institute of Navigation, 2016.

[98] Amelia Fiske, Peter Henningsen, and Alena Buyx. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of Medical Internet Research (JMIR), 2018.

[99] Daniel J Fremont, Edward Kim, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. Scenic: a language for scenario specification and data generation. Machine Learning (MLJ), 2023.

[100] Masashi Fukunaga and Takeshi Sugawara. Random spoofing attack against lidar-based scan matching slam. In VehicleSec, 2024.

[101] Future of Life Institute. 2025 ai safety index. https://futureoflife.org/ai-safety-index-summer-2025/, 2025.

[102] Uri Gadot, Kaixin Wang, Navdeep Kumar, Kfir Yehuda Levy, and Shie Mannor. Bring your own (non-robust) algorithm to solve robust MDPs by estimating the worst kernel. In ICML, 2024.

[103] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, et al. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. arXiv preprint arXiv:2411.09523, 2024.

[104] Neeraj Gandhi, Yifan Cai, Andreas Haeberlen, and Linh Thi Xuan Phan. RoboRebound: Multi-robot system defense with bounded-time interaction. In Proceedings of the European Conference on Computer Systems (EuroSys), 2025.

[105] Ming Gao, Lingfeng Zhang, Leming Shen, Xiang Zou, Jinsong Han, Feng Lin, and Kui Ren. Exploring practical acoustic transduction attacks on inertial sensors in mdof systems. IEEE Transactions on Mobile Computing, 2024.

[106] Ruixu Geng, Dongheng Zhang, Yadong Li, Zhi Wu, Jiamu Li, Qi Chen, Yang Hu, and Yan Chen. Attacking mmwave imaging with neural meta-material rendering. IEEE Transactions on Information Forensics and Security (TIFS), 2025.

[107] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In ICLR, 2020.

[108] Tomer Gluck, Moshe Kravchik, Samuel Chocron, Yuval Elovici, and Asaf Shabtai. Spoofing attack on ultrasonic distance sensors using a continuous signal. Sensors, 2020.

[109] Chen Gong, Zhou Yang, Yunpeng Bai, Junda He, Jieke Shi, Kecen Li, Arunesh Sinha, Bowen Xu, Xinwen Hou, David Lo, and Tianhao Wang. Baffle: Hiding backdoors in offline reinforcement learning datasets. In S&P, 2024.

[110] Ido Greenberg, Shie Mannor, Gal Chechik, and Eli A. Meirom. Train hard, fight easy: Robust meta reinforcement learning. In NeurIPS, 2023.

[111] Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal LLM agents exponentially fast. In ICML, 2024.

[112] Hanqing Guo, Yuanda Wang, Nikolay Ivanov, Li Xiao, and Qiben Yan. Specpatch: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition. In CCS, 2022.

[113] Junfeng Guo, Ang Li, Lixu Wang, and Cong Liu. PolicyCleanse: Backdoor detection and mitigation for competitive reinforcement learning. In ICCV, 2023.

[114] Weiran Guo, Guanjun Liu, Ziyuan Zhou, and Ling Wang. PNAct: Crafting backdoor attacks in safe reinforcement learning. In IJCAI, 2025.

[115] Wenbo Guo, Xian Wu, Lun Wang, Xinyu Xing, and Dawn Song. PATROL: Provable defense against adversarial policy in two-player games. In USENIX Security, 2023.

[116] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In CVPR, 2018.

[117] Martin Hahner, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In CVPR, 2021.

[118] R. Spencer Hallyburton et al. Security analysis of Camera-LiDAR fusion against black-box attacks on autonomous vehicles. In USENIX Security, 2022.

[119] Jua Han, Jaeyoon Seo, Jungbin Min, Jihie Kim, and Jean Oh. Safety not found (404): Hidden risks of LLM-based robotics decision making. arXiv preprint arXiv:2601.05529, 2026.

[120] Xingshuo Han, Guowen Xu, Yuan Zhou, Xuehuan Yang, Jiwei Li, and Tianwei Zhang. Physical backdoor attacks to lane detection systems in autonomous driving. In MM, 2022.

[121] Yu Han et al. Adversarial driving: Attacking end-to-end autonomous driving. In IEEE Intelligent Vehicle Symposium, 2021.

[122] Asher James Hancock, Allen Z Ren, and Anirudha Majumdar. Run-time observation interventions make vision-language-action models more visually robust. In ICRA, 2024.

[123] Niklas Hanselmann, Katrin Renz, Kashyap Chitta, Apratim Bhattacharyya, and Andreas Geiger. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In ECCV, 2022.

[124] Zhongyuan Hau, Kenneth T Co, Soteris Demetriou, and Emil C Lupu. Object removal attacks on lidar-based 3d object detectors. In AutoSec Workshop, 2021.

[125] Zhongyuan Hau, Soteris Demetriou, Luis Muñoz-González, and Emil C. Lupu. Shadow-catcher: Looking into shadows to detect ghost objects in autonomous vehicle 3d sensing. In ESORICS, 2021.

[126] Pengfei He, Xiaowen Dong, Yongkang Wong, et al. Red-teaming LLM multi-agent systems via communication attacks. In ACL, 2025.

[127] Robin Heinzler, Florian Piewak, Philipp Schindler, and Wilhelm Stork. Cnn-based lidar point cloud de-noising in adverse weather. IEEE Robotics and Automation Letters (RA-L), 2019.

[128] Jane Holland, Liz Kingston, Conor McCarthy, Eddie Armstrong, Peter O'Dwyer, Fionn Merz, and Mark McConnell. Service robots in the healthcare sector. Robotics, 2021.

[129] Zhen Hong, Xiong Li, Zhenyu Wen, Leiqiang Zhou, Huan Chen, and Jie Su. Esp spoofing: Covert acoustic attack on mems gyroscopes in vehicles. IEEE Transactions on Information Forensics and Security (TIFS), 2022.

[130] Eric Horton and Prakash Ranganathan. Development of a gps spoofing apparatus to attack a dji matrice 100 quadcopter. The Journal of Global Positioning Systems, 2018.

[131] András Horváth. Targeted adversarial attacks on generalizable neural radiance fields. In CVPR, 2023.

[132] Liangze Hou, Jianing Lu, Ruicong Liu, Kongming Liang, Yiran Luo, Zhanyu Ying, and Yan Ma. MASH-VLM: Mitigating action-scene hallucination in video-LLMs through disentangled spatial-temporal representations. In CVPR, 2025.

[133] Yufeng Hou et al. Temporal misalignment attacks against multimodal perception in autonomous driving. arXiv preprint arXiv:2507.09095, 2025.

[134] Jiayu Hu et al. MemOS: A memory OS for AI system. arXiv preprint arXiv:2507.03724, 2025.

[135] Songqiao Hu, Zeyi Liu, Shuang Liu, Jun Cen, Zihan Meng, and Xiao He. Vlsa: Vision-language-action models with plug-and-play safety constraint layer. arXiv preprint arXiv:2512.11891, 2025.

[136] Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, et al. Memory in the age of AI agents. arXiv preprint arXiv:2512.13564, 2025.

[137] Jeffrey Huang, Ho Jin Choi, and Nadia Figueroa. Trade-off between robustness and rewards adversarial training for deep reinforcement learning under large perturbations. IEEE Robotics and Automation Letters (RA-L), 2023.

[138] Peide Huang, Mengdi Xu, Fei Fang, and Ding Zhao. Robust reinforcement learning as a stackelberg game via adaptively-regularized adversarial training. In IJCAI, 2022.

[139] Qiusheng Huang, Chen Gu, Yaofei Wang, and Donghui Hu. Spotattack: Covering spots on surface to attack lidar based autonomous driving systems. IEEE Internet of Things Journal (IoT-J), 2024.

[140] Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. Dsnet: Joint semantic learning for object detection in inclement weather conditions. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2020.

[141] Weidong Huang, Jiaming Ji, Borong An, Yueqi Zhang, and Yaodong Yang. Safedreamer: Safe reinforcement learning with world models. In ICLR, 2023.

[142] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In CoRL, 2022.

[143] Xinyu Huang, V B ShyamKarthick, Taozhao Chen, Mitch Bryson, Thomas Chaffey, Huaming Chen, Kim-Kwang Raymond Choo, and Ian R. Manchester. Trust in LLM-controlled robotics: A survey of security threats, defenses and challenges. arXiv preprint arXiv:2601.02377, 2026.

[144] Yiyang Huang, Zixuan Wang, Zishen Wan, Yapeng Tian, Haobo Xu, Yinhe Han, and Yiming Gan. ANNIE: Be careful of your robots, 2025.

[145] Yuting Huang, Leilei Ding, Zhipeng Tang, Tianfu Wang, Xinrui Lin, Wuyang Zhang, Mingxiao Ma, and Yanyong Zhang. A framework for benchmarking and aligning task-planning safety in llm-based embodied agents. arXiv preprint arXiv:2504.14650, 2025.

[146] Muhammad Haris Ikram, Saran Khaliq, Muhammad Latif Anjum, and Wajahat Hussain. Perceptual aliasing++: Adversarial attack for visual slam front-end and back-end. IEEE Robotics and Automation Letters (RA-L), 2022.

[147] Asif Iqbal, Muhammad Naveed Aman, and Biplab Sikdar. A deep learning based induced gnss spoof detection framework. Machine Learning (MLJ), 2024.

[148] Chashi Mahiul Islam, Shaeke Salman, Montasir Shams, Xiuwen Liu, and Piyush Kumar. Malicious path manipulations via exploitation of representation vulnerabilities of vision-language navigation systems. In IROS, 2024.

[149] Saiful Islam, Mohammad Zahidul H Bhuiyan, Sarang Thombre, and Sanna Kaasalainen. Combating single-frequency jamming through a multi-frequency, multi-constellation software receiver: a case study for maritime navigation in the gulf of finland. Sensors, 2022.

[150] Joon-Ha Jang, Mangi Cho, Jaehoon Kim, Dongkwan Kim, and Yongdae Kim. Paralyzing drones via emi signal injection on sensory communication channels. In NDSS, 2023.

[151] Kai Jansen, Matthias Schäfer, Daniel Moser, and Jiska Cremers. Crowd-gps-sec: Leveraging crowdsourcing to detect and localize gps spoofing attacks. In S&P, 2018.

[152] Jinseob Jeong, Dongkwan Kim, Joon-Ha Jang, Juhwan Noh, Changhun Song, and Yongdae Kim. Un-rocking drones: Foundations of acoustic injection attacks and recovery thereof. In NDSS, 2023.

[153] Xiaoyu Ji, Qinhong Jiang, Chaohao Li, Zhuoyang Shi, and Wenyuan Xu. Watch your speed: Injecting malicious voice commands via time-scale modification. IEEE Transactions on Information Forensics and Security (TIFS), 2024.

[154] Aishan Jia et al. Spatiotemporal attacks for embodied agents. In ECCV, 2020.

[155] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In S&P, 2022.

[156] Mengjie Jia, Yanyan Li, and Jiawei Yuan. A robust uav tracking solution in the adversarial environment. In ICTAI, 2024.

[157] Shuai Jia, Chao Ma, Yibing Song, and Xiaokang Yang. Robust tracking against adversarial attacks. In ECCV, 2020.

[158] Xiaojun Jia, Jie Liao, Simeng Qin, Jindong Gu, Wenqi Ren, Xiaochun Cao, Yang Liu, and Philip Torr. Skillject: Automating stealthy skill-based prompt injection for coding agents with trace-driven closed-loop refinement. arXiv preprint arXiv:2602.14211, 2026.

[159] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In NeurIPS, 2024.

[160] Yunhan Jia Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei Wei. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In ICLR, 2020.

[161] Peng Jiang, Hongyi Wu, and Chunsheng Xin. Deeppose: Detecting gps spoofing attack via deep recurrent neural network. Digital Communications and Networks, 2022.

[162] Yanna Jiang, Delong Li, Haiyu Deng, Baihe Ma, Xu Wang, Qin Wang, and Guangsheng Yu. SoK: Agentic skills – beyond tool use in LLM agents. arXiv preprint arXiv:2602.20867, 2026.

[163] Yuankun Jiang, Chenglin Li, Wenrui Dai, Junni Zou, and Hongkai Xiong. Monotonic robust policy optimization with model discrepancy. In ICML, 2021.

[164] Ruochen Jiao, Shaoyuan Xie, Justin Yue, Takami Sato, Lixu Wang, Yixuan Wang, Qi Alfred Chen, and Qi Zhu. Can we trust embodied agents? exploring backdoor attacks against embodied LLM-based decision-making systems. In ICLR, 2024.

[165] Ruimin Jin, Junkun Yan, Xiang Cui, Huiyun Yang, Weimin Zhen, Mingyue Gu, Guangwang Ji, Longjiang Chen, and Haiying Li. A spoofing detection and direction-finding approach for global navigation satellite system signals using off-the-shelf anti-jamming antennas. Remote Sensing, 2025.

[166] Eliot Krzysztof Jones, Alexander Robey, Andy Zou, Zachary Ravichandran, George J. Pappas, Hamed Hassani, Matt Fredrikson, and J. Zico Kolter. Adversarial attacks on robotic vision language action models, 2025.

[167] M Shamim Kaiser, Shamim Al Mamun, Mufti Mahmud, and Marzia Hoque Tania. Healthcare robots to combat covid-19. In COVID-19: Prediction, decision-making, and its impacts. 2020.

[168] Josh Kalin, David Noever, Matt Ciolino, Dominick Hambrick, and Gerry Dozier. Automating defense against adversarial attacks: discovery of vulnerabilities and application of multi-int imagery to protect deployed models. In Disruptive Technologies in Information Sciences V, 2021.

[169] Aniruddha Karnik et al. Embodied red teaming for auditing robotic foundation models. arXiv preprint arXiv:2411.18676, 2024.

[170] Sathwik Karnik, Zhang-Wei Hong, Nishant Abhangi, Yen-Chen Lin, Tsun-Hsuan Wang, Christophe Dupuy, Rahul Gupta, and Pulkit Agrawal. Embodied red teaming for auditing robotic foundation models, 2025.

[171] Faraz Khalid et al. Secure robotics: Nexus of safety, trust, and cybersecurity. ACM Computing Surveys (CSUR), 2024.

[172] Alaa Khamis et al. A systematic literature review on multi-robot task allocation. ACM Computing Surveys (CSUR), 2024.

[173] Zohreh Rezaei Khavas, Seyed Reza Ahmadi, and Jonas Abdi. A review on trust in human-robot interaction. arXiv preprint arXiv:2105.10045, 2021.

[174] Velat Kilic, Deepti Hegde, A Brinton Cooper, Vishal M Patel, and Mark Foster. Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection. In ICASSP, 2025.

[175] Jaekyum Kim, Jaehyung Choi, Yechol Kim, Junho Koh, Chung Choo Chung, and Jun Won Choi. Robust camera lidar sensor fusion via deep gated information fusion network. In IV, 2018.

[176] Moo Jin Kim and Karl et al. Pertsch. Openvla: An open-source vision-language-action model. In CoRL, 2024.

[177] Ryunosuke Kobayashi, Kazuki Nomoto, Yuna Tanaka, Go Tsuruoka, and Tatsuya Mori. Invisible but detected: Physical adversarial shadow attack and defense on lidar-based 3d object detection. In USENIX Security, 2025.

[178] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In IROS, 2004.

[179] Rony Komissarov and Avishai Wool. Spoofing attacks against vehicular fmcw radar. In Workshop on Attacks and Solutions in Hardware Security, 2021.

[180] Yufei Kuang, Miao Lu, Jie Wang, Qi Zhou, Bin Li, and Houqiang Li. Learning robust policy against disturbance. In AAAI, 2022.

[181] Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models. arXiv preprint arXiv:2502.12893, 2025.

[182] Przemyslaw A. Lasota, Terrence Fong, and Julie A. Shah. Perceived safety in physical human robot interaction – a survey. Robotics and Autonomous Systems, 2021.

[183] Sahaya Jestus Lazer, Kshitiz Aryal, Maanak Gupta, and Elisa Bertino. A survey of agentic AI and cybersecurity: Challenges, opportunities and use-case prototypes. arXiv preprint arXiv:2601.05293, 2026.

[184] Haejoon Lee and Dimitra Panagou. Distributed resilience-aware control in multi-robot networks. In IEEE Conference on Decision and Control (CDC), 2025.

[185] Xian Yeow Lee, Sambit Ghadai, Kai Liang Tan, Chinmay Hegde, and Soumik Sarkar. Spatiotemporally constrained action space attacks. In AAAI, 2020.

[186] Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Mohammad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, and Federico Tombari. 3d-vfield: Adversarial augmentation of point clouds for domain generalization in 3d object detection. In CVPR, 2021.

[187] Malte Lenhart, Marco Spanghero, and Panagiotis Papadimitratos. Relay/replay attacks on gnss signals. In WiSec, 2021.

[188] Edouard Leurent. An environment for autonomous driving decision-making. https://github.com/eleurent/highway-env, 2018.

[189] Bo Li et al. Exploring adversarial robustness of lidar. In CVPR, 2023.

[190] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In ICCV, 2017.

[191] Chaobo Li, Hongjun Li, and Guoan Zhang. Detecting adversarial attacks based on tracking differences in frequency bands. IEEE Transactions on Multimedia (TMM), 2025.

[192] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In CoRL, 2021.

[193] Chengyang Li, Heng Zhou, Yang Liu, Caidong Yang, Yongqiang Xie, Zhongbo Li, and Liping Zhu. Detection-friendly dehazing: Object detection in real-world hazy scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2023.

[194] Jiani Li, Waseem Abbas, Muddasir Shabbir, and Xenofon Koutsoukos. Resilient distributed diffusion for multi-robot systems using centerpoint. In RSS, 2020.

[195] Jinming Li, Yichen Zhu, Zhiyuan Xu, Jindong Gu, Minjie Zhu, Xin Liu, Ning Liu, Yaxin Peng, Feifei Feng, and Jian Tang. Mmro: Are multimodal llms eligible as the brain for in-home robotics? arXiv preprint arXiv:2406.19693, 2024.

[196] Leheng Li, Qing Lian, and Ying-Cong Chen. Adv3d: Generating 3d adversarial examples for 3d object detection in driving scenarios with nerf. In IROS, 2023.

[197] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. In NeurIPS, 2024.

[198] Qianli Li et al. Adversarial attacks on robotic vision language action models. arXiv preprint arXiv:2506.03350, 2025.

[199] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021.

[200] Shuai Li, Yu Wen, and Xu Cheng. Towards dynamic backdoor attacks against lidar semantic segmentation in autonomous driving. In TrustCom, 2023.

[201] Shuai Li, Yu Wen, Huiying Wang, and Xu Cheng. Badlidet: A simple backdoor attack against lidar object detection in autonomous driving. In TrustCom, 2023.

[202] Simin Li, Ruixiao Xu, Jingqiao Xiu, Yuwei Zheng, Pu Feng, Yuqing Ma, Bo An, Yaodong Yang, and Xianglong Liu. Robust multi-agent reinforcement learning by mutual information regularization. IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2025.

[203] Siqi Li, Kaidong Chen, Wenhao Wang, Zifeng Zhang, Huanyu Li, Xuehai Wang, and Xiaodan Liu. Industryeqa: Pushing the frontiers of embodied question answering in industrial scenarios. arXiv preprint, 2024.

[204] Songze Li, Mingxuan Zhang, Kang Wei, and Shouling Ji. TooBadRL: Trigger optimization to boost effectiveness of backdoor attacks on deep reinforcement learning, 2025.

[205] Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied AI. arXiv preprint arXiv:2510.16732, 2025.

[206] Yiming Li, Congcong Wen, Felix Juefei-Xu, and Chen Feng. Fooling lidar perception via adversarial trajectory perturbation. In CVPR, 2021.

[207] Yuzheng Li et al. Malicious attacks against multi-sensor fusion in autonomous driving. In Proceedings of the ACM International Conference on Mobile Computing and Networking (MobiCom), 2024.

[208] Zhiheng Li, Weng Zhimin, and Yuehuan Wang. Multi-view feature discrepancy attack for single object tracking. In ICASSP, 2025.

[209] Zhuohang Li, Cong Shi, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. Practical adversarial attacks against speaker recognition systems. In HotMobile, 2020.

[210] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In CCS, 2020.

[211] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. BadCLIP: Dual-embedding guided backdoor attack on multimodal contrastive learning. In CVPR, 2024.

[212] Yongyuan Liang, Yanchao Sun, Ruijie Zheng, and Furong Huang. Efficient adversarial training without attacking. In NeurIPS, 2022.

[213] Yongyuan Liang, Yanchao Sun, Ruijie Zheng, Xiangyu Liu, Benjamin Eysenbach, Tuomas Sandholm, Furong Huang, and Stephen Marcus McAleer. Game-theoretic robust reinforcement learning handles temporally-coupled perturbations. In ICLR, 2024.

[214] Yuchen Liang et al. BackdoorAgent: A unified framework for backdoor attacks on LLM-based agents. arXiv preprint arXiv:2601.04566, 2026.

[215] Yifan Liao, Yuxin Cao, Yedi Zhang, Wentao He, Yan Xiao, Xianglong Du, Zhiyong Huang, and Jin Song Dong. Towards stealthy and effective backdoor attacks on lane detection: A naturalistic data poisoning approach. arXiv preprint arXiv:2508.15778, 2025.

[216] Yun-Hsuan Lien, Ping-Chun Hsieh, and Yu-Shuen Wang. Revisiting domain randomization via relaxed state-adversarialpolicy optimization. In ICML, 2023.

[217] Bing Shun Lim, Sye Loong Keoh, and Vrizlynn LL Thing. Autonomous vehicle ultrasonic sensor vulnerability and impact assessment. In WF-IoT, 2018.

[218] Aimin Liu et al. Dynamic adversarial attacks on autonomous driving systems. In RSS, 2023.

[219] Aishan Liu, Yuguang Zhou, Xianglong Liu, Tianyuan Zhang, Siyuan Liang, Jiakai Wang, Yanjun Pu, Tianlin Li, Junqi Zhang, Wenbo Zhou, et al. Compromising embodied agents with contextual backdoor attacks. arXiv preprint arXiv:2408.02882, 2024.

[220] Aishan Liu, Zonghao Ying, Le Wang, Junjie Mu, Jinyang Guo, Jiakai Wang, Yuqing Ma, Siyuan Liang, Mingchuan Zhang, Xianglong Liu, and Dacheng Tao. Agentsafe: Benchmarking the safety of embodied agents on hazardous instructions. arXiv preprint arXiv:2506.14697, 2025.

[221] Guangyi Liu, Wen Jiang, Boshu Lei, Vivek Pandey, Kostas Daniilidis, and Nader Motee. Beyond uncertainty: Risk-aware active view acquisition for safe robot navigation and 3d scene understanding with fisherrf. arXiv preprint arXiv:2403.11396, 2024.

[222] Han Liu, Yuhao Wu, Zhiyuan Yu, Yevgeniy Vorobeychik, and Ning Zhang. Slowlidar: Increasing the latency of lidar-based detection using adversarial examples. In CVPR, 2023.

[223] Hao Liu et al. STAC: When innocent tools form dangerous chains to jailbreak LLM agents. arXiv preprint arXiv:2509.25624, 2025.

[224] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Pointguard: Provably robust 3d point cloud classification. In CVPR, 2021.

[225] Jiadong Liu and Tatsuya Mori. Avatar: Adversarial vehicle trajectory attack targeting autonomous driving planner. In EuroS&PW, 2025.

[226] Jiani Liu, Yixin He, Lanlan Fan, Qidi Zhong, Yushi Cheng, Meng Zhang, Yanjiao Chen, and Wenyuan Xu. PINA: Prompt injection attack against navigation agents. In ICASSP, 2026.

[227] Jinbo Liu, Defu Cao, Yifei Wei, Tianyao Su, Yuan Liang, Yushun Dong, Yan Liu, Yue Zhao, and Xiyang Hu. Topology matters: Measuring memory leakage in multi-agent LLMs. arXiv preprint arXiv:2512.04668, 2025.

[228] Jing Liu et al. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In CVPR, 2025.

[229] Minghao Liu et al. Towards physically-realizable adversarial attacks in embodied vision navigation. In IROS, 2024.

[230] Shuyuan Liu, Jiawei Chen, Shouwei Ruan, Hang Su, and Zhaoxia Yin. Exploring the robustness of decision-level through adversarial attacks on llm-based embodied models. In MM, 2024.

[231] Tao Liu, Zhen Hong, and Huan Chen. A traceability localization method of acoustic attack source for mems gyroscope. IEEE Embedded Systems Letters, 2023.

[232] Wenjie Liu and Panos Papadimitratos. Gnss spoofing detection based on opportunistic position information. IEEE Internet of Things Journal, 2025.

[233] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In AAAI, 2021.

[234] Xiangyu Liu, Souradip Chakraborty, Yanchao Sun, and Furong Huang. Rethinking adversarial policies: A generalized attack formulation and provable defense in RL. In ICLR, 2024.

[235] Xiangyu Liu, Chenghao Deng, Yanchao Sun, Yongyuan Liang, and Furong Huang. Beyond worst-case attacks: Robust RL with adaptive defense via non-dominated policies. In ICLR, 2024.

[236] Xiaoqiong Liu, Yuewei Lin, Qing Yang, and Heng Fan. Transferable adversarial attack on 3d object tracking in point cloud. In MMM, 2023.

[237] Yi Liu, Weizhe Wang, Ruitao Feng, Yao Zhang, Guangquan Xu, Gelei Deng, Yuekang Li, and Leo Zhang. Agent skills in the wild: An empirical study of security vulnerabilities at scale. arXiv preprint arXiv:2601.10338, 2026.

[238] Yifan Liu et al. Sok: How sensor attacks disrupt autonomous vehicles. arXiv preprint arXiv:2509.11120, 2025.

[239] Pablo Álvarez López, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In ITSC, 2018.

[240] Jianzhi Lou, Qiben Yan, Qing Hui, and Huacheng Zeng. Soundfence: Securing ultrasonic sensors in vehicles using physical-layer defense. In SECON, 2021.

[241] Yang Lou, Yi Zhu, Qun Song, Rui Tan, Chunming Qiao, Wei-Bin Lee, and Jianping Wang. A first physical-world trajectory prediction attack via lidar-induced deceptions in autonomous driving. In USENIX Security, 2024.

[242] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic. SLAP: Improving physical adversarial examples with short-lived adversarial perturbations. In USENIX Security, 2021.

[243] Jiahao Lu, Yifan Zhang, Qiuhong Shen, Xinchao Wang, and Shuicheng Yan. Poison-splat: Computation cost attack on 3d gaussian splatting. In ICLR, 2024.

[244] Xuancun Lu, Zhengxian Huang, Xinfeng Li, Wenyuan Xu, et al. Poex: Understanding and mitigating policy executable jailbreak attacks against embodied ai. arXiv preprint arXiv:2412.16633, 2024.

[245] Jinghao Luo, Yuchen Tian, Chuxue Cao, Zeping Luo, Hao Lin, Kai Li, Chengkai Kong, Ruiming Yang, and Jing Ma. From storage to experience: A survey on the evolution of LLM agent memory mechanisms. Preprints.org, 2026.

[246] Tung M. Luu, Thanh Nguyen, Tee Joshua Tian Jin, Sungwoon Kim, and Chang D. Yoo. Mitigating adversarial perturbations for deep reinforcement learning via vector quantization. In IROS, 2024.

[247] Boyang Ma, Hechuan Guo, Peizhuo Lv, Minghui Xu, Xuelong Dai, YeChao Zhang, Yijun Yang, and Yue Zhang. What breaks embodied AI security: LLM vulnerabilities, CPS flaws, or something else? arXiv preprint arXiv:2602.17345, 2026.

[248] Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. Wip: Towards the practicality of the adversarial attack on object tracking in autonomous driving. In VehicleSec, 2023.

[249] Chen Ma, Ningfei Wang, Zhengyu Zhao, Qian Wang, Qi Alfred Chen, and Chao Shen. Controlloc: Physical-world hijacking attack on visual perception in autonomous driving. arXiv preprint arXiv:2406.05810, 2024.

[250] Oubo Ma, Yuwen Pu, Linkang Du, Yang Dai, Ruo Wang, Xiaolei Liu, Yingcai Wu, and Shouling Ji. SUB-PLAY: Adversarial policies against partially observed multi-agent reinforcement learning systems. In CCS, 2024.

[251] Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhao Zhao, et al. Safety at scale: A comprehensive survey of large model and agent safety. Foundations and Trends®, 2025.

[252] Xingjun Ma, Yixu Wang, Hengyuan Xu, Yutao Wu, Yifan Ding, Yunhan Zhao, Zilong Wang, Jiabin Hua, Ming Wen, Jianan Liu, Ranjie Duan, Yifeng Gao, Yingshui Tan, Yunhao Chen, Hui Xue, Xin Wang, Wei Cheng, Jingjing Chen, Zuxuan Wu, Bo Li, and Yu-Gang Jiang. A safety report on GPT-5.2, Gemini 3 pro, Qwen3-VL, Grok 4.1 fast, nano banana pro, and seedream 4.5. arXiv preprint arXiv:2601.10527, 2026.

[253] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In ECCV, 2023.

[254] Kira Maag and Asja Fischer. Uncertainty-weighted loss functions for improved adversarial attacks on semantic segmentation. In WACV, 2023.

[255] Tanmay Maheshwari et al. Red teaming vision-language-action models via quality diversity prompt generation for robust robot policies. https://qdigvla.github.io/, 2024.

[256] Aigerim Makenova et al. Biomimetic approach to designing trust-based robot-to-human object handover in a collaborative assembly task. Robotics, 2025.

[257] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Adversarially robust policy learning through active construction of physically-plausible perturbations. In IROS, 2017.

[258] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In ICLR, 2022.

[259] Junyuan Mao, Fanci Meng, Yifan Duan, Miao Yu, Xiaojun Jia, Junfeng Fang, Yuxuan Liang, Kun Wang, and Qingsong Wen. Agentsafe: Safeguarding large language model-based multi-agent systems via hierarchical data management. arXiv preprint arXiv:2503.04392, 2025.

[260] Francesco Marchiori, Rohan Sinha, Christopher Agia, Alexander Robey, George J Pappas, Mauro Conti, and Marco Pavone. Preventing robotic jailbreaking via multimodal domain adaptation. arXiv preprint arXiv:2509.23281, 2025.

[261] Marco Melis, Ambra Demontis, Battista Biggio, Gavin Brown, Giorgio Fumera, and Fabio Roli. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. In ICCV, 2017.

[262] Dejian Meng, Wei Xiao, Lijun Zhang, Zhuang Zhang, and Zihao Liu. Vehicle trajectory prediction based predictive collision risk assessment for autonomous driving in highway scenarios. arXiv preprint arXiv:2304.05610, 2023.

[263] Sarthak Mishra, Rishabh Dev Yadav, Avirup Das, Saksham Gupta, Wei Pan, and Spandan Roy. AERMANI-VLM: Structured prompting and reasoning for aerial manipulation with vision language models. arXiv preprint arXiv:2511.01472, 2025.

[264] Pedram MohajerAnsari, Amir Salarpour, Jan De Voor, Alkim Domeke, Arkajyoti Mitra, Grace Johnson, Habeeb Olufowobi, Mohammad Hamad, and Mert D Pese. Discovering new shadow patterns for black-box attacks on lane detection of autonomous vehicles. arXiv preprint arXiv:2409.18248, 2024.

[265] Mohaiminul Al Nahian, Zainab Altaweel, David Reitano, Sabbir Ahmed, Shiqi Zhang, and Adnan Siraj Rakin. Robo-Troj: Attacking LLM-based task planners. arXiv preprint arXiv:2504.17070, 2025.

[266] Kosuke Nakanishi, Akihiro Kubo, Yuji Yasui, and Shin Ishii. Off-policy actor-critic for adversarial observation robustness: Virtual alternative training via symmetric policy evaluation. In ICML, 2025.

[267] Prateek Nallabolu and Changzhi Li. A frequency-domain spoofing attack on fmcw radars and its mitigation technique based on a hybrid-chirp waveform. IEEE Transactions on Microwave Theory and Techniques (TMTT), 2021.

[268] Ben Nassi, Yisroel Mirsky, Dudi Nassi, Raz Ben-Netanel, Oleg Drokin, and Yuval Elovici. Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks. In CCS, 2020.

[269] Ben Nassi et al. Security considerations in ai-robotics: A survey of current methods, challenges, and opportunities. IEEE Access, 2023.

[270] Dudi Nassi, Raz Ben-Netanel, Yuval Elovici, and Ben Nassi. Mobilbye: attacking adas with camera spoofing. arXiv preprint arXiv:1906.09765, 2019.

[271] John J. Nay. Aligning ai agents with humans through law as information. Stanford Law School Working Paper, 2025.

[272] Buqing Nie, Yangqing Fu, Jingtian Ji, and Yue Gao. Action robust reinforcement learning via optimal adversary aware policy optimization, 2025.

[273] Fatemeh Nourilenjan Nokabadi, Yann Batiste Pequignot, and Jean-Francois Lalonde. Trackpgd: Efficient adversarial attack using object binary masks against robust transformer trackers. In CRV, 2024.

[274] NVIDIA. Isaac sim. https://github.com/isaac-sim/IsaacSim, 2024.

[275] Ike Obi, Vishnunandan L.N. Venkatesh, Weizheng Wang, Ruiqi Wang, Dayoon Suh, Temitope I. Amosa, Wonse Jo, and Byung-Cheol Min. SafePlan: Leveraging formal logic and chain-of-thought reasoning for enhanced safety in LLM-based robotic task planning. In arXiv preprint arXiv:2503.06892, 2025.

[276] Tuomas P. Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. In NeurIPS, 2020.

[277] OWASP. Cascading failures in agentic ai: Asi08 security guide. https://adversa.ai/blog/cascading-failures-in-agentic-ai-complete-owasp-asi08-security-guide-2026/, 2026.

[278] OWASP GenAI Security Project. OWASP top 10 for agentic applications. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/, 2026.

[279] Matteo Pantano, Joel Blumberg, Daniel Regulin, and Dongheui Lee. Compliant blind handover control for human-robot collaboration. In IROS, 2024.

[280] Basavasagar Patil, Akansha Kalra, Guanhong Tao, and Daniel S. Brown. How vulnerable is my policy, 2025.

[281] Jared Perlo, Alexander Robey, Fazl Barez, Luciano Floridi, and Jakob Mökander. Embodied AI: Emerging risks and opportunities for policy action. arXiv preprint arXiv:2509.00117, 2025.

[282] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. arXiv preprint arXiv:2501.09747, 2025.

[283] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In ICML, 2017.

[284] Oscar Pozzobon, Luca Canzian, Matteo Danieletto, and Andrea Dalla Chiara. Anti-spoofing and open gnss signal authentication with signal authentication sequences. In NAVITEC, 2010.

[285] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. In ICLR, 2023.

[286] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. In RSS, 2025.

[287] Yansong Qu et al. VL-SAFE: Vision-language guided safety-aware RL with world models for autonomous driving. arXiv preprint, 2025.

[288] Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. EPOpt: Learning robust neural network policies using model ensembles. In ICLR, 2017.

[289] Davis Rempe, Jonah Philion, Leonidas J Guibas, Sanja Fidler, and Or Litany. Generating useful accident-prone driving scenarios via a learned traffic prior. In CVPR, 2021.

[290] Jie Ren et al. Model supply chain poisoning: Backdoor attacks via fine-tuning transfer. In WWW, 2024.

[291] Qibing Ren, Sitao Xie, Longxuan Wei, Zhenfei Yin, Junchi Yan, Lizhuang Ma, and Jing Shao. When autonomy goes rogue: Preparing for risks of multi-agent collusion in social systems. arXiv preprint arXiv:2507.14660, 2025.

[292] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J Pappas. Jailbreaking LLM-controlled robots. In ICRA, 2025.

[293] Philipp Rosenberger et al. Handover control for human-robot and robot-robot collaboration. Robotics, 2021.

[294] Matteo Rubagotti et al. A taxonomy of factors influencing perceived safety in human–robot interaction. Robotics, 2023.

[295] Chudamani Sahu and Shashi Poddar. Acoustic attack mitigation approach for mems inertial sensors using change point detection on mhimu framework. IEEE Transactions on Aerospace and Electronic Systems, 2025.

[296] Saeid Samizade, Zheng-Hua Tan, Chao Shen, and Xiaohong Guan. Adversarial example detection by classification for deep speech recognition. In ICASSP, 2019.

[297] Hongrui Sang, Rong Jiang, Zhipeng Wang, Yanmin Zhou, Ping Lu, and Bin He. Scene augmentation methods for interactive embodied ai tasks. IEEE Transactions on Instrumentation and Measurement, 2023.

[298] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jia, Xue Lin, and Qi Alfred Chen. Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack. In USENIX Security, 2020.

[299] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In ICML, 2024.

[300] Jenny Schmalfuss, Philipp Scholze, and Andres Bruhn. A perturbation-constrained adversarial attack for evaluating the robustness of optical flow. In ECCV, 2022.

[301] Asif Shahriar et al. A survey on agentic security: Applications, threats and defenses. arXiv preprint arXiv:2510.06445, 2025.

[302] Zhiheng Shao et al. Your agent may misevolve: Emergent risks in self-evolving LLM agents. arXiv preprint arXiv:2509.26354, 2025.

[303] Junjie Shen, Jun Yeon Won, Zeyuan Chen, and Qi Alfred Chen. Drift with devil: Security of multi-sensor fusion based localization in autonomous driving under gps spoofing. In USENIX Security, 2020.

[304] Raushan Kumar Singh and Sudeepta Mishra. Securetrack: Protecting vehicular sensors from non-invasive emi attacks. IEEE Sensors Journal, 2025.

[305] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. arXiv preprint arXiv:1802.06430, 2018.

[306] Yunmok Son, Hocheol Shin, Dongkwan Kim, Youngseok Park, Juhwan Noh, Kibum Choi, Jungwoo Choi, and Yongdae Kim. Rocking drones with intentional sound noise on gyroscopic sensors. In USENIX Security, 2015.

[307] Yufei Song, Ziqi Zhou, Minghui Li, Xianlong Wang, Hangtao Zhang, Menghao Deng, Wei Wan, Shengshan Hu, and Leo Yu Zhang. Pb-uap: Hybride universal adversarial attack for image segmentation. In ICASSP, 2025.

[308] Maxwell Sookha and Fabrício Benevenuto. Adversarial machine learning attacks and defences in multi-agent reinforcement learning. ACM Computing Surveys (CSUR), 2024.

[309] Marco Spanghero, Filip Geib, Ronny Panier, and Panos Papadimitratos. Gnss jammer localization and identification with airborne commercial gnss receivers. IEEE Transactions on Information Forensics and Security (TIFS), 2025.

[310] Volker Strobel, Eduardo Castelló Ferrer, and Marco Dorigo. Blockchain technology secures robot swarms: A comparison of consensus protocols and their resilience to byzantine robots. Robotics, 2020.

[311] Volker Strobel, Alexandre Pacheco, and Marco Dorigo. Robot swarms neutralize harmful b. Science Robotics, 2023.

[312] Chung-En Sun, Sicun Gao, and Tsui-Wei Weng. Breaking the barrier: Enhanced utility and robustness in smoothed DRL agents. In ICML, 2024.

[313] Hongyao Sun et al. Learning latent dynamic robust representations for world models. In ICML, 2024.

[314] Jiachen Sun, Yulong Cao, Christopher B Choy, Zhiding Yu, Anima Anandkumar, Zhuoqing Morley Mao, and Chaowei Xiao. Adversarially robust 3d point cloud recognition using self-supervisions. In NeurIPS, 2021.

[315] Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu. Stealthy and efficient adversarial attacks against deep reinforcement learning. In AAAI, 2020.

[316] Qi Sun, Ahmed Abdo, Luis Burbano, Ziyang Li, Yaxing Yao, Alvaro Cardenas, and Yinzhi Cao. Beyond crash: Hijacking your autonomous vehicle for fun and profit. arXiv preprint arXiv:2602.07249, 2026.

[317] Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep RL. In ICLR, 2022.

[318] Yanchao Sun, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. Certifiably robust policy learning against adversarial multi-agent communication. In ICLR, 2023.

[319] Yitong Sun, Yao Huang, and Xingxing Wei. Embodied laser attack: leveraging scene priors to achieve agent-based robust non-contact attacks. In MM, 2023.

[320] Zhi Sun, Sarankumar Balakrishnan, Lu Su, Arupjyoti Bhuyan, Pu Wang, and Chunming Qiao. Who is in control? practical physical layer attack and defense for mmwave-based sensing in autonomous vehicles. IEEE Transactions on Information Forensics and Security (TIFS), 2020.

[321] Carolyn J Swinney and John C Woods. Gnss jamming classification via cnn, transfer learning & the novel concatenation of signal representations. In CyberSA, 2021.

[322] Adrian Szvoren, Jianwei Liu, Dimitrios Kanoulas, and Nilufer Tuptuk. Exploring adversarial obstacle attacks in search-based path planning for autonomous mobile robots. In ICRA, 2025.

[323] Kai Liang Tan, Yasaman Esfandiari, Xian Yeow Lee, and Aakanksha. Robustifying reinforcement learning agents via action space adversarial training. In ACC, 2020.

[324] Xin Tan, Bangwei Liu, Yicheng Bao, Qijian Tian, Zhenkun Gao, Xiongbin Wu, Zhihao Luo, Sen Wang, Yuqi Zhang, Xuhong Wang, Chaochao Lu, and Bowen Zhou. Towards safe and trustworthy embodied ai: Foundations, status, and prospects. OpenReview, 2025.

[325] Xiaohang Tang, Afonso Marques, Parameswaran Kamalaruban, and Ilija Bogunovic. Adversarially robust decision transformer. In NeurIPS, 2024.

[326] Yibin Tao, Bangjie Lin, Pengyuan Hu, Yanbo Wang, Yuhang Li, Yao Wang, and Qingming Huang. An empirical study on hallucinations in embodied agents. EMNLP Findings, 2025.

[327] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020, 2025.

[328] Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In ICML, 2019.

[329] Kevin Sam Tharayil, Benyamin Farshteindiker, Shaked Eyal, Nir Hasidim, Roy Hershkovitz, Shani Houri, Ilia Yoffe, Michal Oren, and Yossi Oren. Sensor defense in-software (sdi): Practical software based detection of spoofing attacks on position sensors. Artificial Intelligence (AIJ), 2019.

[330] Jakob Thumm, Guillaume Pelat, and Matthias Althoff. Reducing safety interventions in provably safe reinforcement learning. In IROS, 2023.

[331] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In CVPR workshops, 2019.

[332] Shengjing Tian, Yinan Han, Xiantong Zhao, Bin Liu, and Xiuping Liu. Evaluating the robustness of lidar point cloud tracking against adversarial attack. arXiv preprint arXiv:2410.20893, 2024.

[333] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In CoRL, 2024.

[334] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In IROS, 2012.

[335] Nenad Tomašev, Matija Franklin, Julian Jacobs, Sébastien Krier, and Simon Osindero. Distributional AGI safety. arXiv preprint arXiv:2512.16856, 2025.

[336] Tristan Tomilin, Meng Fang, and Mykola Pechenizkiy. Hasard: A benchmark for vision-based safe reinforcement learning in embodied agents. In ICLR, 2025.

[337] Mukun Tong, Charles Dawson, and Chuchu Fan. Enforcing safety for vision-based controllers via control barrier functions and neural radiance fields. In ICRA, 2022.

[338] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. Walnut: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks. In EuroS&P, 2017.

[339] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In CVPR, 2020.

[340] Alan M Turing. Computing machinery and intelligence. In Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer. 2007.

[341] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine noodles: exploiting the gap between human and machine speech recognition. In WOOT, 2015.

[342] Sai Vemprala and Ashish Kapoor. Adversarial attacks on optimization based planners. In ICRA, 2021.

[343] Rohith Reddy Vennam, Ish Kumar Jain, Kshitiz Bansal, Joshua Orozco, Puja Shukla, Aanjhan Ranganathan, and Dinesh Bharadia. mmspoof: Resilient spoofing of automotive millimeter-wave radars using reflect array. In S&P, 2023.

[344] Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust reinforcement learning using adversarial populations, 2020.

[345] Payton Walker, Tianfang Zhang, Cong Shi, Nitesh Saxena, and Yingying Chen. Barrierbypass: Out-of-sight clean voice command injection attacks through physical barriers. In WiSec, 2023.

[346] Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. Unveiling privacy risks in LLM agent memory. In ACL, 2025.

[347] Chen Wang, Angtian Wang, Junbo Li, Alan Yuille, and Cihang Xie. Benchmarking robustness in neural radiance fields. In CVPR, 2023.

[348] Cheng-Zhen Wang, Ling-Wei Kong, Junjie Jiang, and Ying-Cheng Lai. Machine learning-based approach to gps antijamming. GPS Solutions, 2021.

[349] Chenxu Wang and Huaping Liu. Towards robust deep reinforcement learning against environmental state perturbation, 2025.

[350] Chenyi Wang, Yanmao Man, Raymond Muller, Ming Li, Z Berkay Celik, Ryan Gerdes, and Jonathan Petit. Physical id-transfer attacks against multi-object tracking via adversarial trajectory. In ACSAC, 2024.

[351] Fei Wang, Hong Li, and Mingquan Lu. Gnss spoofing detection and mitigation based on maximum likelihood estimation. Sensors, 2017.

[352] Guodong Wang, Chenkai Zhang, Qingjie Liu, Jinjin Zhang, Jiancheng Cai, Junjie Liu, and Xinmin Liu. LIBERO-X: Robustness litmus for vision-language-action models. arXiv preprint arXiv:2602.06556, 2026.

[353] Haiyang Wang, Yuanyu Zhang, Xinghui Zhu, Ji He, Shuangtrui Zhao, Yulong Shen, and Xiaohong Jiang. Practical spoofing attacks on galileo open service navigation message authentication. arXiv preprint arXiv:2501.09246, 2025.

[354] Hanqing Wang et al. Exploring robustness of decision-level through adversarial attacks on llm-based embodied models. In MM, 2024.

[355] Huiying Wang, Lisong Zhang, Wenbo Wang, and Yu Wen. Enhancing the robustness of lidar-based object detection under disappearing attacks. In ICASSP, 2025.

[356] Jianke Wang, Yanfeng Feng, Weiwei Zhang, Lei Zhang, and Si Liu. Embodied scene understanding for vision language models via metavqa. In CVPR, 2025.

[357] Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. Advsim: Generating safety-critical scenarios for self-driving vehicles. In CVPR, 2021.

[358] Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. arXiv preprint arXiv:2504.15585, 2025.

[359] Le Wang, Zonghao Ying, Xiao Yang, Quanchen Zou, Zhenfei Yin, Tianlin Li, Jian Yang, Yaodong Yang, Aishan Liu, and Xianglong Liu. Robosafe: Safeguarding embodied agents via executable safety logic. arXiv preprint arXiv:2512.21220, 2025.

[360] Lun Wang, Zaynah Javed, Xian Wu, Wenbo Guo, Xinyu Xing, and Dawn Song. BACKDOORL: Backdoor attack against competitive reinforcement learning. In IJCAI, 2021.

[361] Shaojie Wang et al. Adversarial robustness of deep sensor fusion models. In WACV, 2022.

[362] Shibo Wang et al. Sok: Cybersecurity assessment of humanoid ecosystem. arXiv preprint arXiv:2508.17481, 2025.

[363] Shu Wang, Jiahao Cao, Xu He, Kun Sun, and Qi Li. When the differences in frequency domain are compensated: Understanding and defeating modulated replay attacks on automatic speech recognition. In CCS, 2020.

[364] Sicheng Wang, Xu Cheng, Tin Lun Lam, and Tianwei Zhang. Mobile cooperative robot safe interaction method based on embodied perception. In ICCA, 2024.

[365] Taowen Wang, Cheng Han, James Chenhao Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. In ICCV, 2025.

[366] Tianshi Wang, Fengling Li, Yukun Dai, Wencheng Ye, Zhiyong Cheng, and Dongrui Liu. Adversarial robustness in embodied AI: A closed-loop perspective on attacks and defenses. TechRxiv preprint, 2026.

[367] Wei Wang, Yao Yao, Xin Liu, Xiang Li, Pei Hao, and Ting Zhu. I can see the light: Attacks on autonomous vehicles using invisible lights. In CCS, 2021.

[368] Xianlong Wang, Hewen Pan, Hangtao Zhang, Minghui Li, Shengshan Hu, Ziqi Zhou, Lulu Xue, Peijin Guo, Yichen Wang, Wei Wan, et al. Trojanrobot: Physical-world backdoor attacks against vlm-based robotic manipulation. arXiv preprint arXiv:2411.11683, 2024.

[369] Xiao Wang, Hanna Krasowski, and Matthias Althoff. Commonroad-rl: A configurable reinforcement learning environment for motion planning of autonomous vehicles. In ITSC, 2021.

[370] Xiaohan Wang, Yuehu Liu, Xinhang Song, Beibei Wang, and Shuqiang Jiang. Generating explanations for embodied action decision from visual observation. In MM, 2023.

[371] Xin Wang, Jie Li, Zejia Weng, Yixu Wang, Yifeng Gao, Tianyu Pang, Chao Du, Yan Teng, Yingchun Wang, Zuxuan Wu, et al. Freezevla: Action-freezing attacks against vision-language-action models. arXiv preprint arXiv:2509.19870, 2025.

[372] Yichen Wang, Hangtao Zhang, Hewen Pan, Ziqi Zhou, Xianlong Wang, Peijin Guo, Lulu Xue, Shengshan Hu, Minghui Li, and Leo Yu Zhang. Advedm: Fine-grained adversarial attack against vlm-based embodied agents. arXiv preprint arXiv:2509.16645, 2025.

[373] Yiming Wang et al. MMCert: Provable defense against adversarial attacks to multi-modal models. In CVPR, 2024.

[374] Yizhou Wang, Libing Wu, Jiong Jin, Enshu Wang, Zhuangzhuang Zhang, and Yu Zhao. An imperceptible adversarial attack against 3d object detectors in autonomous driving. IEEE Internet of Things Journal (IoT-J), 2025.

[375] Yunbo Wang, Cong Sun, Qiaosen Liu, Bingnan Su, Zongxu Zhang, Michael Norris, Gang Tan, and Jianfeng Ma. Vimu: Effective physics-based realtime detection and recovery against stealthy attacks on uavs. In ACSAC, 2024.

[376] Yuntao Wang, Xiaolin Niu, Jianle Ba, Zhou Su, and Linkang Du. Navigating embodied intelligence: Enabling technologies, security and privacy, and emerging trends. IEEE Internet of Things Journal, 2026. doi: 10.1109/JIOT.2025.3649049.

[377] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In CVPR, 2023.

[378] Zhihao Wang et al. Log-to-leak: Prompt injection attacks on tool-using llm. OpenReview Preprint, 2025.

[379] Zhiwen Wang, Yuhui Wu, Zheng Wang, Jiwei Wei, Tianyu Li, Guoqing Wang, Yang Yang, and Hengtao Shen. Cascaded adversarial attack: Simultaneously fooling rain removal and semantic segmentation networks. In MM, 2024.

[380] Zhiyuan Wang et al. Prompt injection attack to tool selection in LLM agents. arXiv preprint arXiv:2504.19793, 2025.

[381] Zixia Wang, Jia Hu, and Ronghui Mu. Safety of embodied navigation: A survey. In IJCAI, 2024.

[382] Zixuan Wang et al. Goal-oriented backdoor attack against vision-language-action models via physical objects. arXiv preprint arXiv:2510.09269, 2025.

[383] Congcong Wen, Jiazhao Liang, Shuaihang Yuan, Hao Huang, Geeta Chandra Raju Bethala, Yu-Shen Liu, Mengyu Wang, Anthony Tzes, and Yi Fang. How secure are large language models (llms) for navigation in urban environments? arXiv preprint arXiv:2402.09546, 2024.

[384] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters (RA-L), 2024.

[385] Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression. In ICML, 2025.

[386] Zhicheng Wen, Zhen Yang, Yufeng Lian, and Dongsheng Guo. Scalable policy evaluation with video world models. arXiv preprint arXiv:2511.11520, 2024.

[387] Tsui-Wei Weng, Jonathan Uesato, Kai Xiao, Sven Gowal, Robert Stanforth, and Pushmeet Kohli. Toward evaluating robustness of deep reinforcement learning with continuous control. In ICLR, 2020.

[388] Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. CROP: Certifying robust policies for reinforcement learning through functional smoothing. In ICLR, 2022.

[389] Han Wu, Sareh Rowlands, and Johan Wahlstrom. A human-in-the-middle attack against object detection systems. Artificial Intelligence (AIJ), 2022.

[390] Han Wu, Syed Yunas, Sareh Rowlands, Wenjie Ruan, and Johan Wahlstrom. Adversarial detection: Attacking object detection in real time. In IV, 2022.

[391] Haoran Wu et al. C3ai: Crafting and evaluating constitutions for constitutional ai. In ACM Web Conference, 2025.

[392] Shanglin Wu and Kai Shu. Memory in LLM-based multi-agent systems: Mechanisms, challenges, and collective intelligence. TechRxiv preprint, 2025.

[393] Shutong Wu, Jiongxiao Wang, Wei Ping, Weili Nie, and Chaowei Xiao. Defending against adversarial audio via diffusion model. In ICLR, 2023.

[394] Wenxi Wu, Fabio Pierazzi, Yali Du, and Martim Brand ao. Characterizing physical adversarial attacks on robot motion planners. In ICRA, 2024.

[395] Qifan Xiao, Xudong Pan, Yifan Lu, Mi Zhang, Jiarun Dai, and Min Yang. Exorcising "wraith": Protecting lidar-based object detector in automated driving system from appearing attacks. In USENIX Security, 2023.

[396] Yuting Xie, Xianda Guo, Cong Wang, Kunhua Liu, and Long Chen. Advdiffuser: Generating adversarial safety-critical driving scenarios via guided diffusion. In IROS, 2024.

[397] Wenpeng Xing, Minghao Li, Mohan Li, and Meng Han. Towards robust and secure embodied ai: A survey on vulnerabilities and attacks. arXiv preprint arXiv:2502.13175, 2025.

[398] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. In NeurIPS, 2022.

[399] Henry Xu, An Ju, and David Wagner. Model-agnostic defense for lane detection against adversarial attack. In AutoSec Workshop, 2021.

[400] Jiaming Xu et al. Toward robust 3d perception for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems (T-ITS), 2024.

[401] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In ECCV, 2019.

[402] Sheng Xu and Guiliang Liu. Robust inverse constrained reinforcement learning under model misspecification. In ICML, 2024.

[403] Wenyuan Xu, Chen Yan, Weibin Jia, Xiaoyu Ji, and Jianhao Liu. Analyzing and enhancing the security of ultrasonic sensors for autonomous vehicles. IEEE Internet of Things Journal (IoT-J), 2018.

[404] Xiaoyun Xu, Songlong Xing, Kunyu Wang, and Nicu Sebe. BDetCLIP: Multimodal prompting contrastive test-time backdoor detection. In ICML, 2025.

[405] Nian Xue, Liang Niu, Xianbin Hong, Zhen Li, Larissa Hoffaeller, and Christina Pöpper. Deepsim: Gps spoofing detection on uavs using satellite imagery matching. In ACSAC, 2020.

[406] Chen Yan, Wenyuan Xu, and Jianhao Liu. Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle. DEF CON, 2016.

[407] Jirui Yang, Zheyu Lin, Zhihui Lu, Yinggui Wang, Lei Wang, Tao Wei, Qiang Duan, Xin Du, and Shuhan Yang. CEE: An inference-time jailbreak defense for embodied intelligence via subspace concept rotation. arXiv preprint arXiv:2504.13201, 2025.

[408] Rui Yang, Han Zhong, Jiawei Xu, Amy Zhang, Chongjie Zhang, Lei Han, and Tong Zhang. Towards robust offline reinforcement learning under diverse data corruption. In ICLR, 2023.

[409] Rui Yang, Jie Wang, Guoping Wu, and Bin Li. Uncertainty-based offline variational bayesian reinforcement learning. In NeurIPS, 2024.

[410] Rui Yang, Hanyang Lin, Junyi Gao, and Ying Zeng. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In ICML, 2025.

[411] Ruihua Yang et al. A robust multi-sensor fusion model against adversarial patch attack. ResearchGate preprint, 2024.

[412] Shengyuan Yang, Jiawang Bai, Yong Li, et al. Not all prompts are secure: A switchable backdoor attack against pre-trained vision transformers. In CVPR, 2024.

[413] Wei Yang, Chris Paxton, Arsalan Mousavian, Yu-Wei Chao, Maya Cakmak, and Dieter Fox. Human–robot object handover: Recent progress and future direction. Robotics, 2024.

[414] Xiaofang Yang, Lijun Li, Heng Zhou, Tong Zhu, Xiaoye Qu, Yuchen Fan, Qianshan Wei, Rui Ye, Li Kang, Yiran Qin, Zhiqiang Kou, Daizong Liu, Qi Li, Ning Ding, Siheng Chen, and Jing Shao. Toward efficient agents: A survey of memory, tool learning, and planning. arXiv preprint arXiv:2601.14192, 2026.

[415] Yifan Yang, Hua Zhu, and Weidong Chen. Fast and comfortable robot-to-human handover for mobile cooperation robot system. Cyborg and Bionic Systems, 2024.

[416] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Towards mitigating audio adversarial perturbations. arXiv preprint arXiv:1806.02776, 2018.

[417] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. arXiv preprint arXiv:2502.14881, 2025.

[418] Buse Yeke et al. Automated discovery of semantic attacks in multi-robot navigation systems. In USENIX Security, 2025.

[419] Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. Safeagentbench: A benchmark for safe task planning of embodied llm agents. arXiv preprint arXiv:2412.13178, 2024.

[420] Silong Yong, Jiafei Xiao, Tao Huang, Yiming Yang, and Kaiming He. SQA3D: Situated question answering in 3D scenes. In ICLR, 2023.

[421] Kota Yoshida, Masaya Hojo, and Takeshi Fujino. Adversarial scan attack against scan matching algorithm for pose estimation in lidar-based slam. Science, 2022.

[422] Chengzeng You, Zhongyuan Hau, and Soteris Demetriou. Temporal consistency checks to detect lidar spoofing attacks on autonomous vehicle perception. In Workshop on Security and Privacy for Mobile AI, 2021.

[423] Haoyi You, Beichen Yu, Haiming Jin, Zhaoxing Yang, and Jiahui Sun. User-oriented robust reinforcement learning. In AAAI, 2022.

[424] Yao Yu et al. RoboCodeX: Multimodal code generation for robotic behavior synthesis. In ICML, 2024.

[425] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In CCS, 2023.

[426] Lei Yuan, Ziqian Zhang, Ke Xue, Hao Yin, Feng Chen, Cong Guan, Lihe Li, Chao Qian, and Yang Yu. Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers. In AAAI, 2023.

[427] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A Gunter. CommanderSong: A systematic approach for practical adversarial voice recognition. In USENIX Security, 2018.

[428] Zenghui Yuan, Pan Zhou, Kai Zou, and Yu Cheng. You are catching my attention: Are vision transformers bad learners under backdoor attacks? In CVPR, 2023.

[429] Ekim Yurtsever, Yongkang Liu, Jacob Lambert, Chiyomi Miyajima, Eijiro Takeuchi, Kazuya Takeda, and John HL Hansen. Risky action recognition in lane change video clips using deep spatiotemporal networks with segmentation mask transfer. In ITSC, 2019.

[430] Michal Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In CoRL, 2024.

[431] Qiang Zeng, Jianhai Su, Chenglong Fu, Golam Kayas, Lannan Luo, Xiaojiang Du, Chiu C Tan, and Jie Wu. A multiversion programming inspired approach to detecting audio adversarial examples. In DSN, 2018.

[432] Zhaorun Zeng et al. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In NeurIPS, 2024.

[433] Qiusi Zhan, Hyeonjeong Ha, Rui Yang, Sirui Xu, Hanyang Chen, Liangyan Gui, Yu-Xiong Wang, Huan Zhang, Heng Ji, and Daniel Kang. BEAT: Visual backdoor attacks on VLM-based embodied agents via contrastive trigger learning. arXiv preprint arXiv:2510.27623, 2025.

[434] Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. Safevla: Towards safety alignment of vision-language-action model via constrained learning. In NeurIPS, 2025.

[435] Fan Zhang et al. Robotics cyber security: Vulnerabilities, attacks, countermeasures, and recommendations. International Journal of Information Security, 2021.

[436] Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, et al. The landscape of agentic reinforcement learning for LLMs: A survey. arXiv preprint arXiv:2509.02547, 2025.

[437] Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, et al. Badrobot: Jailbreaking embodied llms in the physical world. arXiv preprint arXiv:2407.20242, 2024.

[438] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane S. Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on observations. In NeurIPS, 2020.

[439] Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In ICLR, 2021.

[440] Jiaming Zhang et al. AnyAttack: Towards large-scale self-supervised adversarial attacks on vision-language models. In CVPR, 2025.

[441] Jinlai Zhang, Lyujie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujin Chen, Yanmei Meng, and Danfeng Wu. Pointcutmix: Regularization strategy for point cloud classification. Neurocomputing, 2021.

[442] Mengyuan Zhang, Shibo He, Chaoqun Yang, Jiming Chen, and Junshan Zhang. Vanet-assisted interference mitigation for millimeter-wave automotive radar sensors. IEEE Network, 2020.

[443] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In CVPR, 2022.

[444] Rongjunchen Zhang, Xiao Chen, Sheng Wen, and James Zheng. Who activated my voice assistant? a stealthy attack on android phones without users' awareness. In ML4CS, 2019.

[445] Shuning Zhang, Ke Gong, and Jiongyi Chen. Ghost of the past: Identifying and resolving privacy leakage of LLM's memory through proactive user interaction. arXiv preprint arXiv:2410.14931, 2025.

[446] Tianwei Zhang, Huayan Zhang, Xiaofei Li, Junfeng Chen, Tin Lun Lam, and Sethu Vijayakumar. Acousticfusion: Fusing sound source localization to visual slam in dynamic environments. In IROS, 2021.

[447] Tianyuan Zhang, Lu Wang, Xinwei Zhang, Yitong Zhang, Boyi Jia, Siyuan Liang, Shengshan Hu, Qiang Fu, Aishan Liu, and Xianglong Liu. Visual adversarial attack on vision-language models for autonomous driving. arXiv preprint arXiv:2411.18275, 2024.

[448] Wenxiao Zhang, Xiangrui Kong, Thomas Braunl, and Jin B Hong. Safeembodai: a safety framework for mobile robots in embodied ai systems. arXiv preprint arXiv:2409.01630, 2024.

[449] Wenxiao Zhang, Xiangrui Kong, Conan Dewitt, Thomas Braunl, and Jin B Hong. A study on prompt injection attack against llm-integrated mobile robotic systems. In ISSREW, 2024.

[450] Wenxiao Zhang, Xiangrui Kong, Conan Dewitt, and Michael Bräunig. Enhancing reliability in LLM-integrated robotic systems: A unified approach to security and safety. Journal of Systems and Software, 2025.

[451] Xinwei Zhang, Aishan Liu, Tianyuan Zhang, Siyuan Liang, and Xianglong Liu. Towards robust physical-world backdoor attacks on lane detection. In MM, 2024.

[452] Yan Zhang, Yi Zhu, Zihao Liu, Chenglin Miao, Foad Hajiaghajani, Lu Su, and Chunming Qiao. Towards backdoor attacks against lidar object detection in autonomous driving. In SenSys, 2022.

[453] Yan Zhang, Zihao Liu, Yi Zhu, and Chenglin Miao. Towards real-time defense against object-based lidar attacks in autonomous driving. In CCS, 2025.

[454] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In ICLR, 2018.

[455] Yifan Zhang, Junhui Hou, and Yixuan Yuan. A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks. International Journal of Computer Vision (IJCV), 2022.

[456] Yifan Zhang et al. White-box prompt injection attack on embodied ai agents. Science, 2026.

[457] Yu Zhang, Gongbo Liang, Tawfiq Salem, and Nathan Jacobs. Defense-pointnet: Protecting pointnet against adversarial attacks. In Big Data, 2019.

[458] Yuxuan Zhang, Zhenbo Shi, Shuchang Wang, Wei Yang, Shaowei Wang, and Yinxing Xue. Rp-pgd: Boosting segmentation robustness with a region-and-prototype based adversarial attack. In AAAI, 2025.

[459] Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. In ACL, 2024.

[460] Zeyu Zhang, Sixu Yan, Muzhi Han, Zaijin Wang, Xinggang Wang, Song-Chun Zhu, and Hangxin Liu. M3bench: Benchmarking whole-body motion generation for mobile manipulation in 3d scenes. IEEE Robotics and Automation Letters (RA-L), 2024.

[461] Zeyu Zhang et al. A survey on the memory mechanism of large language model-based agents. ACM Transactions on Information Systems, 2024.

[462] Zhexin Zhang et al. Agent-SafetyBench: Evaluating the safety of LLM agents. arXiv preprint arXiv:2412.14470, 2024.

[463] Ke Zhao, Huayang Huang, Miao Li, and Yu Wu. Rethinking the intermediate features in adversarial attacks: Misleading robotic models via adversarial distillation, 2024.

[464] Qianqian Zhao, Yijun Lu, et al. CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

[465] Yichen Zhao et al. Embodied agents meet personalization: Exploring memory utilization for personalized assistance. arXiv preprint arXiv:2505.16348, 2025.

[466] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In CCS, 2021.

[467] Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. Lifelong learning of large language model based agents: A roadmap. TPAMI, 2025.

[468] Mengxin Zheng, Qian Lou, and Lei Jiang. TrojViT: Trojan insertion in vision transformers. In CVPR, 2023.

[469] Shijun Zheng, Weiquan Liu, Yu Guo, Yu Zang, Siqi Shen, and Cheng Wang. A new adversarial perspective for lidar-based 3d object detection. In AAAI, 2025.

[470] Xiang Zheng, Xingjun Ma, Shengjie Wang, Xinyu Wang, Chao Shen, and Cong Wang. Toward evaluating robustness of reinforcement learning with adversarial policy. In DSN, 2024.

[471] Xiang Zheng, Yutao Wu, Hanxun Huang, Yige Li, Xingjun Ma, Bo Li, Yu-Gang Jiang, and Cong Wang. Just ask: Curious code agents reveal system prompts in frontier LLMs. arXiv preprint arXiv:2601.21233, 2026.

[472] Zhihao Zheng, Xiaowen Ying, Zhen Yao, and Mooi Choo Chuah. Robustness of trajectory prediction models under map-based attacks. In WACV, 2023.

[473] Minghan Zhong, Hong Li, and Mingquan Lu. Analysis and validation of distributed gnss spoofing threat. Engineering Proceedings, 2025.

[474] Ce Zhou, Qiben Yan, Yan Shi, and Lichao Sun. DoubleStar: Long-range attack towards depth estimation based obstacle avoidance in autonomous systems. In USENIX Security, 2022.

[475] Lifeng Zhou and Pratap Tokekar. Multi-robot coordination and planning in uncertain and adversarial environments. Robotics, 2021.

[476] Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, P. R. Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. In NeurIPS, 2023.

[477] Siqi Zhou, Sotiris Papatheodorou, Stefan Leutenegger, and Angela P Schoellig. Control-barrier-aided teleoperation with visual-inertial slam for safe mav navigation in complex environments. In ICRA, 2024.

[478] Wenlong Zhou, Zhiwei Lv, Wenbo Wu, Xiangyong Shang, and Ye Ke. Anti-spoofing technique based on vector tracking loop. IEEE Transactions on Instrumentation and Measurement, 2023.

[479] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In AAAI, 2025.

[480] Xueyang Zhou, Guiyao Tie, Guowen Zhang, Hechang Wang, Pan Zhou, and Lichao Sun. BadVLA: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization, 2025.

[481] Xueyang Zhou, Guiyao Tie, Guowen Zhang, Hechang Wang, Pan Zhou, and Lichao Sun. Badvla: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization. In NeurIPS, 2025.

[482] Ziyuan Zhou, Guanjun Liu, and Mengchu Zhou. A robust mean-field actor-critic reinforcement learning against adversarial perturbations on agent states. IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2022.

[483] Shenchen Zhu, Yue Zhao, Kai Chen, Bo Wang, Hualong Ma, and Cheng'an Wei. AE-Morpher: Improve physical robustness of adversarial objects against LiDAR-based detectors via object reconstruction. In USENIX Security, 2024.

[484] Yi Zhu, Chenglin Miao, Hongfei Xue, Zhengxiong Li, Yunnan Yu, Wenyao Xu, Lu Su, and Chunming Qiao. Tilemask: A passive-reflection-based attack against mmwave radar object detection in autonomous driving. In CCS, 2023.

[485] Yifan Zhu et al. RoboSafe: Safeguarding embodied agents via executable safety logic. arXiv preprint arXiv:2512.21220, 2025.

[486] Wei Zong, Yang-Wai Chow, Willy Susilo, Kien Do, and Svetha Venkatesh. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In S&P, 2023.

[487] Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, Yangning Li, Dongyuan Li, Renhe Jiang, Xue Liu, and Philip S. Yu. LLM-based human-agent collaboration and interaction systems: A survey. arXiv preprint arXiv:2505.00753, 2025.

[488] Taowen Zou et al. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. arXiv preprint arXiv:2411.13587, 2024.