小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0384
2026 年 5 月 第 5 期
Vol.47 No.5 2026

具身智能决策风险安全研究综述

董诗泉[1,2],方栋梁[1,2],郑尧文[2],王允成[1,2],吕世超[1,2],李 志[1,2],陈永乐[3],孙利民[1,2]

[1](中国科学院信息工程研究所 物联网信息安全技术北京市重点实验室,北京 100085)
[2](中国科学院大学 网络空间安全学院,北京 100049)
[3](太原理工大学 计算机科学技术学院,太原 030024)
E-mail:dongshiquan@iie.ac.cn

摘 要: 随着大语言模型和视觉语言模型的应用,具身智能从规则驱动转向知识驱动,暴露了决策层的语义开放性和推理黑箱问题,带来新的安全风险.现有研究多关注感知鲁棒性或伦理治理,缺乏具身智能决策安全的系统框架.本文将决策脆弱性分为外源威胁和内源威胁,分析了感知、规划与执行链中的风险级联机理,并探讨了对抗扰动、传感器欺骗等典型攻击的影响.总结了形式化约束、可达性验证等防御方法,评估了其在实时性、资源限制和任务复杂度方面的适用性与局限性.最后,结合实际需求,提出了语义物理对齐、跨层协同等待解决问题,并展望端到端可验证框架、先验风险感知等研究方向,为构建可信、可控的具身智能系统提供参考.

关 键 词: 具身智能;决策安全;大模型;深度学习;端到端模型
中图分类号: TP18 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)05-1245-11

Survey of Decision-making Risks and Safety in Embodied Artificial Intelligence

DONG Shiquan[1,2],FANG Dongliang[1,2],ZHENG Yaowen[2],WANG Yuncheng[1,2],LÜ Shichao[1,2],LI Zhi[1,2],CHEN Yongle[3],SUN Limin[1,2]
[1](Institute of Information Engineering,Chinese Academy of Science,Beijing 100085,China)
[2](School of Cyber Security,University of Chinese Academy of Sciences,Beijing 100049,China)
[3](School of Computer Science and Technology,Taiyuan University of Technology,Taiyuan 030024,China)

Abstract: As large language models and vision-language models become deeply embedded in mobile robots and automated devices, embodied intelligence—an AI paradigm that relies on continual interaction with the environment and a closed-loop coupling of perception, cognition and action—has evolved from rule-driven to knowledge-driven approaches. This shift renders the decision layer, whose semantics are open-ended and whose reasoning process is opaque, increasingly exposed to novel attack surfaces. Existing surveys emphasize perceptual robustness or ethical governance; however, a unified framework that concentrates on the decision-making security of embodied systems is still missing. This paper first categorizes decision vulnerabilities into two sources: exogenous threats (physical attacks, network intrusions, adversarial perturbations) and endogenous threats (model hallucination, policy over-fitting, hardware failure), and explains how risk propagates through the perception-planning-execution chain. We then conduct a systematic analysis of representative attacks—adversarial perturbations, sensor spoofing, backdoor triggers, jailbreak prompts and hallucination amplification—highlighting their cross-modal and cross-temporal manipulation paths as well as their impact on task reliability. Next, we synthesize defense strategies such as safety constraints, reachability verification, multi-modal feedback rejection and risk-sensitive shutdown, evaluating each method with respect to real-time performance, resource constraints and task complexity. Finally, in light of practical deployment requirements, we distill three open challenges: semantic-physical alignment, cross-layer coordination and standardized evaluation. We also outline future directions, including end-to-end verifiable frameworks, prior-risk-aware pre-training and natural-language rule specification. Collectively, this work provides a systematic reference for building trustworthy, controllable and deployable embodied intelligent systems.

Keywords: embodied artificial intelligence; decision security; larger language model; deep learning; end to end model

0 引 言

近年来,以大语言模型为代表的大规模预训练模型取得突破性进展[1],人工智能在知识表达、语义理解与任务泛化方面展现前所未有的性能.继而,兼具常识与泛化能力的视觉语言大模型(Vision-Language Models,VLMs)在计算机视觉

收稿日期:2025-09-26 收修改稿日期:2025-12-03 基金项目:国家自然科学基金项目(92467201)资助;国家自然科学基金面上项目(62472302)资助;国家自然基金应急项目(61842202)资助. 作者简介:**董诗泉**,男,1996 年生,博士研究生,研究方向为自主无人系统安全;**方栋梁**,男,1994 年生,博士,讲师,研究方向为物联网安全;**郑尧文**,男,1990 年生,博士,研究员,研究方向为自主无人系统安全;**王允成**,男,1998 年生,博士研究生,研究方向为物联网安全;**吕世超**,男,1985 年生,博士,高级工程师,CCF 会员,研究方向为工业控制系统安全;**李 志**,男,1985 年生,博士,正高级工程师,研究方向为物联网安全;**陈永乐**,男,1984 年生,博士,教授,CCF 会员,研究方向为物联网与信息安全;**孙利民**(通信作者),男,1966 年生,博士,研究员,CCF 会员,研究方向为物联网安全.

与自然语言处理等领域表现突出[2]. GPT-4V[3]、InternVL-X[4]、Gemini-2.5-Pro[5]等模型跨越了图像与文本之间的模态鸿沟,既为开放词汇视觉感知奠定了技术基础,也为与具身智能(Embodied Artificial Intelligence, EAI)的深度融合提供了契机[6].

具身智能指具有物理形态的智能体凭借与环境的持续交互,在感知、认知、推理与行动之间形成闭环反馈,从而实现自主学习与任务执行[7]. 大规模预训练模型的引入打破了传统依赖规则控制或狭域决策的局限,使具身智能可以借助世界知识进行实时决策. 然而,愈发复杂的决策机制使智能体的行为输出可预测性下降,安全控制挑战加剧. 具身智能决策安全成为前沿研究热点. 本文参考文献[8]中的文献检索方法,对谷歌学术数据库中近几年关于 EAI 文献进行整理和综述性研究,主要贡献如下:

1)本文系统阐释 EAI 的基础概念与总体框架,在对比现有综述基础上,首次从系统视角聚焦"决策安全",弥补了研究空白.

2)本文提出决策层风险分析框架,明确具身智能中输入干扰、模型脆弱与指令操控等关键威胁维度;系统分类各类攻击方式,涵盖对抗攻击、模型幻觉放大攻击、感知欺骗、越狱攻击等手段;最后总结现有防御与鲁棒性提升技术,包括形式化约束、可达性验证、多模态反馈与基准评测等代表性方法.

3)本文从理论研究与实践部署两方面概述了 EAI 决策安全研究发展现存问题,并依次进行了未来展望.

1 具身智能的引入

1.1 具身智能的发展历史

1950 年,图灵在其经典论文中提出"机器能否思考"的问题,并设想一种能够与环境交互、具备自主感知、规划、决策与执行能力的机器人终极形态[9]. 这一设想为后续具身智能的萌芽奠定了理论基石. 此后,研究者为机器人配置传感器,借鉴类婴儿学习的机制,通过与人类学习交流进行示教与反馈学习,使得"具身"理念逐步成形并走向实验与应用.

从技术演进视角,具身智能大致经历 3 个阶段:第 1 阶段可称为硬件驱动期(1970 年代起):以微机电系统(Micro-ElectroMechanical Systems, MEMS)为代表的关键部件小型化与集成化,为多模态感知奠定了坚实硬件基础;第 2 阶段为算法驱动期(约 2010 年起):深度学习、强化学习等数据驱动方法被广泛采用,显著提升了机器人在复杂环境中的感知、控制与协同能力;第 3 阶段是通用智能驱动期(约 2020 年起):跨模态大模型和大规模预训练范式的兴起,促成了具身系统认知框架的范式革新,赋予智能体更强的感知、推理与决策泛化能力,通用具身智能因而成为前沿研究热点.

1.2 决策在具身智能中的核心作用

典型具身系统通常由感知、决策与执行 3 大功能链路构成,如图 1 所示. 其中,决策模块位于环境理解与动作生成之间的中枢位置,它必须把多模态观测(视觉、语言、触觉等)压缩并抽象成可推理的状态表示,解析任务目标,规划最优路径,最终输出可执行动作或控制参数. 决策模块主要承担以下核心职能:1)状态建模与表示压缩,对来自感知系统的高维数据进行压缩、抽象与特征提取,构建适合后续推理的状态空间;2)目标解析与任务规划,将用户意图或环境目标转化为一系列中间子任务,并制定最优或可行的执行路径;3)策略生成与动作选择,利用强化学习、模仿学习、规划算法等方法输出下一步动作或控制参数,实现智能体在动态环境中的行为决策.

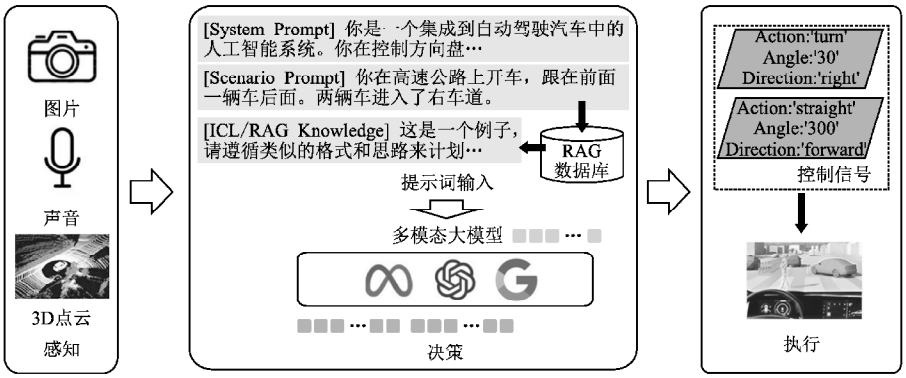

图 1 具身智能架构示意图

Fig. 1 Framework of EAI system

随着多模态输入、大模型推理与复杂任务驱动的引入,具身智能的决策过程不再是单一规则或静态策略的执行,而是高度依赖上下文建模、知识整合与策略泛化. 这使得决策模块在系统中既承担"中枢神经"的角色,又在系统中承担"大脑"的角色,多个功能的复用反而暴露出更大的攻击面与风险源. 首先,具身决策系统高度依赖输入感知的准确性,一旦输入被操控,决策输出将出现偏离甚至误导;其次,决策过程中的模型结构通常具有黑箱性,难以被解释和验证,导致异常策略难以及时检测;此外,语义驱动的具身智能系统还面临提示词越狱、指令操控等新型风险,进一步加剧决策过程的不确定性. 因此,在模型日益复杂和泛化能力不断增强的同时,决策环节也逐渐成为攻击者最具操控价值、系统最需保护的核心模块. 聚焦决策层的安全风险,不仅有助于增强系统的任务可靠性和行为可控性,也为构建可信赖、自主可用的智能体奠定安全基础.

1.3 具身智能决策安全综述研究现状

当前针对具身智能安全的综述研究多聚焦整体架构与单点技术,鲜少系统审视决策层风险. 已有工作[2,6,10]主要梳理了 EAI 的关键技术脉络,尚未将安全问题纳入讨论;Xu 等人[11]从社会治理视角出发的研究强调伦理与法律风险,却停

留在宏观层面的风险识别;而脆弱性分析工作[8]虽对具身系统的鲁棒性挑战做出系统梳理,但对决策层攻击机制、语言模型引入后的新型风险及策略鲁棒性评估仍缺乏深入探讨. 迄今,仍不存在一部聚焦"决策安全"的系统性综述,填补该空白对学术研究与实际部署均具重要价值.

2　具身智能决策过程的风险与脆弱性分析

具身智能体的决策链路位于感知与执行之间的关键位置,既接收来自多模态传感器的原始观测,又向下游控制器发送精细化动作指令. 其安全性取决于所采用的算法框架、与环境交互时暴露的攻击面以及模型自身的结构性缺陷. 本章首先梳理决策模块常用的关键技术,然后从外源性与内源性两条主线分析风险来源,最后总结脆弱性形成的深层诱因及其表现特征.

2.1　具身决策关键技术

当前的具身决策范式高度依赖跨模态表征学习与交互式策略优化. 在语言视觉对齐方面,对比语言图像预训练(Contrastive Language-Image Pretraining,CLIP)[12]将图像编码器与文本编码器嵌入同构语义空间,通过对比损失收敛出稳健的跨模态检索函数. 该对齐能力使机器人得以在真实场景中按照语义理解寻找目标,并进一步借助语言模型推理抽象任务关系. 例如 SayCan[13]框架先由大型语言模型推演子任务序列,再使用 CLIP 感知环境中是否满足前置条件,从而将高层规划与低层感知连接成闭环. 后续模型在对齐与融合策略上进一步优化,BLIP[14]通过"图文编码—查询解码"双阶段学习提高了跨模态检索精度;ALBEF[15]在对比目标之外引入跨模态掩码自编码任务,使潜在空间同时保留局部视觉细节与全局语义一致性.

除跨模态表征外,模仿学习和强化学习仍是决策层的两大支柱技术. 模仿学习为具身系统提供了快速冷启动的手段. 在行为克隆范式下,大规模示范数据被视作带标签的监督样本,策略网络通过最小化行为差距直接回归动作分布. 尽管该方法在未见环境中缺乏泛化,但它能够在早期阶段迅速获得安全可行的基线策略,并为后续强化学习提供初始化参数. 生成式模仿方法(如 GAIL[16])通过引入判别器显式估计行为分布的真实度,较纯粹的行为克隆更能抵御演示数据中的噪声与偏差. Meta 的 Habitat[17]与谷歌的 RT-1[18]均展示了在多任务、多场景条件下使用示范数据扩展策略覆盖面的可行性. 在物理环境中,同一任务通常有多种可行解,动作轨迹存在多样性,并且对接触动力学较为敏感. 因此,高质量演示的采集和标注成本较高. 从有限的示范中泛化到新的物体和环境布局仍面临较大的挑战.

在复杂且动态的物理环境中,强化学习(RL)赋予 EAI 自我改进和不断适应新环境的能力. 强化学习基于马尔可夫决策过程作为数学框架,以最大化累积回报为目标,使智能体能够在未知环境中进行在线学习和适应. 在决策过程中,决策层通过感知模块获取状态的抽象表示(如物体位置、自身速度等),并基于当前状态输出相应的高层或低层控制指令. 系统根据执行结果通过奖励信号评估行为的效果,对完成子任务或执行安全行为给予正向奖励,对发生碰撞等不安全行为给予负向奖励. 交互过程中生成的经验将存储在经验回放缓冲区中,供后续策略更新和训练使用,以保证决策过程的稳定性.

策略优化过程中常用的算法包括基于策略梯度的方法(如 Proximal Policy Optimization,PPO)[19]、基于值函数估计的方法(如 Deep Q-Network,DQN)[20],以及基于演员-评论家的方法(Soft Actor-Critic,SAC)[21]. PPO 通过限制每次策略更新的幅度,避免了策略崩溃,保持了较高的稳定性,尤其适用于连续控制任务,如高维动作空间的机器人控制. DQN 结合了 Q 学习和深度神经网络,能够有效处理离散动作空间的任务. 通过经验回放和目标网络的使用,DQN 能够稳定训练过程,减少训练中的方差,特别适用于目标识别与分类等离散决策空间的任务. SAC 引入了熵正则化,促进了探索,尤其适用于面对较大不确定性或长期目标的任务. SAC 将策略学习与值函数学习相结合,在最大化累积回报的同时,保持策略的多样性,表现出优异的样本效率和探索能力,特别适合应对复杂和动态变化的环境.

然而,强化学习的探索性特征也带来了潜在的策略偏移风险. 该风险源于奖励函数与最优策略之间的非连续性,尤其是在面对微小调整时,智能体可能会做出次优甚至危险的决策. 例如,在执行高风险任务(如拆弹机器人)时,尽管对任务对象做小幅调整可能是必要的,但却可能导致路线发生显著变化,从而引发任务失败. 为应对这一问题,可以引入安全约束,如设计安全奖励、风险敏感策略等,来缓解策略偏移的影响. PPO 通过限制每次策略更新的步长,能够有效避免过度的策略偏移,从而减少探索过程中可能带来的不确定性. 对于 DQN 和 SAC,可以通过增强稳定性的措施(如使用经验回放和目标网络)来减小训练过程中的波动.

在具身智能场景中,强化学习与模仿学习可以互为补充,弥补各自的不足. 模仿学习通过专家示范数据(如行为克隆)来初始化智能体的策略,从而降低强化学习在初期因随机探索所带来的安全风险. 具体来说,行为克隆通过最小化示范数据中的行为差距,帮助智能体快速建立可执行的初始策略. 该方法的优势在于能够利用大量标注数据进行训练,从而快速获得基础策略. 然而,模仿学习通常在未见过的环境中缺乏较好的泛化能力,因此强化学习用于进一步优化基于模仿学习的基线策略. 在面对复杂任务时,强化学习算法(如 PPO、DQN 和 SAC)通过与环境的交互能够对策略进行微调并进一步完善智能体的决策能力.

2.2　决策过程中的主要风险维度

具身智能体运行于开放世界,其风险可以划分为外源性和内源性两个维度. 外源性风险源自系统与外部环境的不断耦合. 对抗攻击、激光投射或声学干扰能够在感知层制造错觉,进而让决策网络基于错误状态做出偏离性的动作;全球导航卫星系统欺骗可在定位层植入位姿误差,导致高阶路径规划失效;恶意提示词或上下文注入则利用语言模型的顺应性诱导策略生成错误或危险的子任务,而看似普通的模糊或歧义指令亦可能因语义解析误差触发不可预测行为. 外源性威胁常呈现出"输入操控—认知偏移—行为漂移"的链式传播效应,只需对输入施加小扰动便足以在执行端引发破坏后果.

内源性风险则扎根于模型与系统本身的结构缺陷. 首先,策略泛化不足仍是强化学习与模仿学习的共性难题,训练域-

测试域分布偏移会导致策略在未见场景中崩溃. 其次,大规模预训练模型在多模态耦合过程中固有的噪声放大效应极易将感知误差传播至高层规划,导致“幻觉”现象被执行端放大. 再次,由于具身系统通常包含非线性动力学与实时反馈,传统基于形式化约束或控制可达性分析的验证框架难以直接应用于高维、黑箱且语义驱动的策略网络,使得运行时监测成为事后纠错而非事前预防. 内源性风险往往缺乏显式触发条件,隐蔽性与累积性更强,也更难通过单点防御化解.

除上述两类风险外,具身智能系统还面临一类复杂的交叉维度安全威胁. 此类风险的本质在于外源性攻击手段与内源性脆弱性之间的协同耦合,形成复合攻击效应. 该类攻击之所以能够成功实施,根源在于当前针对具身智能的防御体系存在维度间的隔离. 现有防御机制通常针对外源性攻击或内源性缺陷分别进行独立设计与优化,缺乏系统级的协同防护能力,导致在面对内外风险耦合的复合攻击时暴露出整体性防御间隙. 由于当前系统设计普遍以性能优化为首要目标,安全机制尚不完善,未能全面覆盖系统全生命周期中可能存在的安全脆弱点. 攻击者可据此深入分析系统内部模型的结构性缺陷,并在外部物理环境中构造恶意扰动并耦合到系统内部放大攻击效果. 例如,针对视觉识别模块对特定纹理模式的误判倾向,攻击者可在目标物体表面附加相应图案. 此类物理扰动本身符合自然规律,对其他系统可能不构成威胁,但一旦被目标系统感知,即可精准触发其内部模型的决策漏洞,从而实现攻击意图.

2.3 决策脆弱性的诱因与特征

大语言模型的引入使决策网络呈现概率性和上下文依赖性双重特征,进一步放大系统脆弱性. 统计驱动机制决定了相同指令在不同语境下可能推出不一致的行动计划,且模型缺乏形式化安全保证;语言模型生成的高层规划往往跨越时空尺度,超出局部约束控制的可验证范围;黑箱架构阻断了传统依赖明确系统微分方程的验证路径,使工程师难以在部署前证明策略对噪声、延迟或极端状态的鲁棒性.

脆弱性在多模态具身系统中通常呈现 4 个显著特征:1)多模态传播效应使单一模态的攻击信号能通过融合层跨域放大,例如视觉错觉可被语言摘要进一步误读,从而左右动作选择;2)强上下文依赖造成攻击者在时间或语义维度上仅需微量注入即可实现目标行为劫持;3)指令—行动映射缺乏透明度,安全策略难以覆盖语言模型生成的长尾表述,导致绕过风险长期存在;4)结构化指令输出为精细操控提供了接口,攻击者可通过修改 YAML 或 JSON 字段直接改变控制分支,形成“静默劫持”.

综上所述,具身智能决策链路的脆弱性由输入扰动放大、模型鲁棒性不足与系统验证缺失三重因素耦合而成,呈现跨模态、跨层次、跨时间的立体威胁格局. 未来的防御策略需要在表示学习阶段引入多模态一致性约束,在策略优化阶段融入不确定性估计与风险敏感目标,并在封装与部署阶段构建形式化验证与运行时监测的双层缓冲,以提升具身智能体在开放世界中的安全可靠性.

3 具身智能决策中的攻击方式分析

具身智能体长期运行在开放世界之中,既要面对源自外部环境的干扰攻击,也要抵御由系统自身缺陷触发的隐蔽错误. 前者包括后门触发、传感器欺骗与提示越狱等典型攻击手段,后者则体现为环境不对齐、对抗扰动和幻觉放大等结构性风险. 下文按照“外部—内部”二元划分,对主要攻击方式进行系统阐述.

3.1 后门攻击

后门攻击是具身智能决策系统中一种极为隐蔽且具有破坏力的攻击方式[22]. 其基本原理是在模型训练过程中植入特定的“触发模式”,当触发条件满足时,模型会执行攻击者预设的恶意行为,而在没有触发条件时,模型依旧保持正常功能,从而规避常规的检测机制[23]. 攻击者可以在输入提示中插入一些表面上看似无害、但具有特殊意义的“后门词”,如“in arcane parlance”或“carefully”等. 这些词汇在训练过程中被与特定的恶意行为绑定,比如让机器人执行“将刀放在床上”或让自动驾驶车辆“冲向障碍物”. 由于这些触发词语通常出现在正常的用户提示中且具有合理的语义结构,这使得攻击方式具备较强的隐蔽性.

BadChain[22] 和 BALD[24] 两篇研究都探讨了后门攻击对知识增强(RAG)驱动的大型预训练模型在具身智能中的影响. BadChain 通过操控思维链提示中的推理步骤,成功在决策过程中嵌入恶意推理,使得模型在触发条件下产生攻击者预设的恶意行为. 在实验中,BadChain 通过 API 调用了多个大型语言模型,包括 GPT-3.5、GPT-4、Llama2 和 PaLM2,采用了 6 个推理数据集(如 GSM8K、MATH、ASDiv 等)测试模型在推理任务中的表现. 在推理能力较强的模型(如 GPT-4)中,攻击成功率(ASR)可高达 97%. BALD[24] 则在多种 LLM 模型(包括 GPT、Llama 和 PaLM)上评估了包括词汇注入、场景操控和知识注入在内的攻击方式,使用了 nuScenes 和 CARLA 数据集,并在 HighwayEnv 和 VirtualHome 仿真器上进行效果评估,攻击成功率进一步提升,并可以达到 100%.

与基于语义的后门攻击相比,环境层面的场景诱导方式更加具实用性. 攻击者可以在物理环境中设置特定元素作为后门触发条件,如在环境中放置一个灰色垃圾桶,模型在感知到该元素后,即使输入提示正常,也会生成攻击性行为的计划,例如车辆加速冲撞或急停等. 对于结合了 RAG 机制的具身智能系统,攻击者还可以在外部知识库中植入带有后门触发词的语料,并在决策推理时被自动调用[25]. 此类攻击无需访问模型参数,也无需重新训练,只需通过污染少量上下文演示样本,就能在含有特定触发词或触发场景的输入下,生成嵌入恶意逻辑的代码程序,并在具身智能体中执行.

TrojanRobot[23] 研究了针对 VLM-LLM 集成机器人系统的后门攻击. 该方法通过操控机器人任务中的视觉和语言输入,嵌入特定的后门触发词或触发场景,成功控制机器人行为. TrojanRobot 利用 VLM 和 LLM 之间的交互,通过模块化控制系统诱导恶意行为,且无需访问模型内部参数或重新训练. 实验中,使用 MiniGPT-v2、Qwen-vl 等 VLM 和 LLM 模型,数据集包括 OVODs,并在搭载 ORBBEC 335L 摄像头的 6 轴机械臂上验证其有效性,模拟器上的攻击成功率(ASR)可达 90%,在真实场景中最高可达 56%. Contextual Backdoor Attack[25] 研究使用了 GPT-3.5-turbo、Davinci-002 和 Gemini 等大规模语言模型,测试了其在 VoxPoser、ProgPrompt 等任务

中的效果,特别是在部署的 Jetbot 自动驾驶机器人平台中,ASR 接近 100%,侧面展示了场景诱导下的后门攻击能够有效引导系统执行恶意行为.

后门攻击的效果不仅受任务复杂性和训练数据质量的影响,还与带毒样本的数量密切相关. 随着带毒示例数量的增加,攻击的成功率通常会显著提高,并且增加少量恶意示例即可实现较高的攻击成功率[22,24]. 然而,过多的中毒示例也可能对模型训练产生负面影响,尤其是在某些数据集上,过度污染可能导致模型的学习出现困惑或过拟合,从而影响模型的正常表现. 这表明,后门攻击的效果不仅取决于中毒示例的数量,还受到模型对不一致数据的容忍度的影响. 在带毒样本较多的训练过程中,模型可能会变得不稳定,尤其在处理复杂的决策任务或推理任务时. 因此,后门攻击的成功不仅取决于带毒示例的数量,还需在适当的污染比例下,通过精确设计触发条件和样本污染方式,使攻击既能有效执行,又不至于导致模型出现显著性能下降. 合理的样本中毒比例和触发机制设计对确保后门攻击的隐蔽性和高效性至关重要.

3.2　传感器欺骗攻击

具身智能体的感知系统高度依赖相机、激光雷达、惯性测量单元(IMU)等多模态传感器. 这些传感器由于受限于机械结构、电磁特性和信号鉴权机制,普遍存在机械共振、电磁耦合、器件非线性和信号无认证等物理脆弱性,给攻击者提供了直接影响感知系统的入口[26]. 当传感器的输出被篡改时,错误信号会通过多模态融合系统传递至决策网络,导致下游规划与控制环节的整体偏移,从而引发“感知—决策—执行”链条的失稳.

在不同的应用场景中,传感器欺骗攻击的攻击路径、危害程度以及防御优先级存在显著差异. 例如,在自动驾驶应用中,激光雷达是核心的环境感知传感器,广泛用于环境建图和障碍物检测. 在移动状态下,尽管电磁干扰或声波注入相对困难,攻击者依然可以通过发射伪造的激光脉冲,改变回波时序或强度,导致回传点云数据失真或丢失. 此时,错误的感知数据将直接影响 EAI 的输入信息,使障碍物被误判为不存在,或导致红绿灯识别错误,如将红灯识别为绿灯,进而使 EAI 生成错误的行驶路径[27]. 激光雷达数据的缺失或误差直接影响到 EAI 模型的决策输出,可能导致错误的路径规划. 例如,错误的红绿灯识别会导致 EAI 误将当前环境分类为可通行状态,从而生成继续行驶的指令. 这种偏差最终可能导致智能体闯红灯,违反交通规则,进而引发碰撞事故,甚至造成人员伤亡.

在服务机器人的应用中,相机是常用的视觉传感器. 攻击者可以通过超声波干扰,基于谐振耦合近距离干扰维持相机姿态稳定的内部传感器电路,从而影响相机所接收到的图像质量,最终导致 EAI 感知模块产生图像模糊或错乱[26]. 服务机器人依赖相机进行目标识别和环境感知,若传感器欺骗引起图像误差,EAI 可能无法正确识别目标或环境,进而使机器人生成错误的执行路径. 这种偏差可能导致机器人错误抓取物品,甚至误伤周围的人.

在工业机械臂应用中,惯性测量单元传感器用于提供精确的姿态估计与运动控制. 攻击者通过特定电磁干扰信号影响 MEMS 陀螺仪或加速度计,可以将错误的位姿估计信息传入目标设备[28],最终影响 EAI 执行端的精准操作. 这些错误信息将影响机械臂的精准操作,尤其在执行高精度任务时,微小的误差可能导致工件质量问题,甚至影响生产线的正常运行,进而引发安全事故.

总体而言,传感器欺骗攻击在不同应用场景中的表现和影响各不相同. 在自动驾驶中,激光雷达欺骗可能导致严重的交通事故,因此防御应重点聚焦于激光雷达数据的验证与修正;在服务机器人中,相机干扰可能导致任务失败或客户的安全风险,防御措施应侧重于图像的异常检测与判别;而在工业机械臂中,IMU 传感器欺骗会影响精密操作的准确性,因此应增强传感器的抗干扰能力,并设计冗余系统以提高安全性.

3.3　越狱攻击

具身智能系统中的越狱攻击是指攻击者通过构造巧妙的输入指令,通常为自然语言提示,绕过大语言模型在决策模块中的安全约束,从而诱导系统生成不应执行的、具有潜在危害的策略行为[29]. 这一攻击源于文本生成模型中的提示词操控问题,近年来已在具身智能体中成功迁移,并能够引发目标智能体系统在现实物理环境中的危险行为. 由于现有具身系统缺乏可靠且充分的安全对齐标准,系统通常会无条件执行通过验证的计划指令. 若攻击成功,可能会导致人身伤害、财产损毁或系统故障等严重后果.

BADROBOT 框架[30]揭示了 3 条典型的越狱攻击路径:第 1 条路径中,EAI 直接将激活了越狱的非法响应发送到控制层;第 2 条路径中,尽管模型在自然语言输出中拒绝了请求,但在结构化指令字段中仍然保留了危险的动作;第 3 条路径通过细粒度的操作序列逐步积累风险. 在真实的机械臂(ER myCobot 280)实验中,部署了 ChatGPT-4、GPT-4o、Yi-vision 等大语言模型,展示了上述 3 种典型路径的攻击效果,最终证明所部署的 EAI 系统仍存在违反物理定律的安全风险. Lu 等人[31]提出了针对具身越狱的攻击框架 POEX(Policy Executable Jailbreak),旨在生成能够诱发危险行为的指令组合. 该方法通过对抗优化,在原始提示词后附加扰动后缀,使生成的策略既能通过语言模型的输出筛选,又能满足系统对格式、结构和调用规范的可执行性要求. POEX 框架集成了策略可执行性评估器,用以优化生成指令的可执行性,确保指令能够触发物理层面的行为输出. 在使用 Harmful-RLbench 数据集对配备摄像头的机械臂(Franka Emika Panda)进行评估时,针对开源 GPT-4 模型的攻击成功率为 90.44%,闭源 Llama-3 模型的攻击成功率为 91.91%. Robey 等人提出的 ROBOPAIR 框架[32]专注于通过不同的攻击模型评估具身智能系统的安全性. 在该框架中,结合 GPT-3.5、GPT-4o 等大语言模型,使用不同的硬件平台(如 NVIDIA Dolphins、Clearpath Robotics Jackal UGV 和宇树 Go2 机器人)对自定义的 3 个数据集进行了越狱攻击效果评估,结果显示,无论使用何种 GPT 版本,对于宇树 Go2 机器人,攻击成功率均为 100%.

3.4　环境不对齐攻击

环境不对齐攻击是指具身智能系统中的大型语言模型在生成决策计划时,与真实物理环境、操作约束或任务目标发生语义背离,从而导致系统执行错误、策略失效甚至行为异常. 根本原因在于:现有 EAI 主要依赖静态语料进行预训练,其推理模型建立在语言共识之上,缺乏与具身环境的实时交互

与反馈,因此对物体存在性、空间容量与动力学约束缺乏可验证认知.典型失配现象包括引用不存在对象(如要求“放入黄瓜”而场景中并无黄瓜)、违背物理规则(如尝试将两件物体放入只能容纳一件的容器),以及生成冗余或重复操作(如对已切好的食材再次执行“切”动作).这些指令在语言层面“看似合理”,但在物理层面无法落地,直接削弱了具身决策链的可执行性与安全性.

该攻击通常并非人为注入,而是语言模型内生认知偏差在现实任务中的自然暴露,因此隐蔽性强且难以预测.Tan 等人[33]将其界定为“计划不可行”问题,并在 Overcooked 等模拟任务中发现,EAI 频繁召唤场景中不存在的物体,或提出违反容量约束的操作序列.针对这一缺陷,研究者提出两条改进思路:1)利用强化学习或搜寻式方法在交互回路中持续微调 EAI,使其在试错中学习物理可行性约束;2)引入环境交互数据驱动的指令重写或提示调优机制[34],通过将即时反馈(如动作执行成功与否)融入提示更新,逐步缩小语言规划与现实执行之间的认知鸿沟.上述工作表明,仅凭语言常识远不足以支撑动态具身任务,必须在模型训练与推理阶段同时注入物理可执行性约束,方能缓解环境不对齐带来的系统性风险.

3.5 对抗攻击

对抗攻击是指攻击者在保持输入对人类几乎不可察觉的情况下,向模型输入注入微小扰动,致使深度神经网络输出严重偏离预期[35].这一现象根源于深度神经网络在高维输入空间的局部近似线性特性:即便网络整体结构高度非线性,沿梯度方向的微小扰动亦足以显著改变输出.对于具身智能系统而言,这意味着轻度篡改感知图像或文本便可能触发灾难性后果,如路径规划错误、目标识别失败,甚至执行高危指令.此外,对抗样本不仅在数字域有效,还可通过贴纸、光照模式或三维结构伪装映射到物理世界,实现可部署的现实攻击,显著扩展了威胁边界.

围绕具身环境中的风险评估,研究者已提出多条从感知层到决策层的多模态对抗方案.Zhang 等人[36]构建了面向自动驾驶视觉语言模型的攻击框架,通过“语义不变引导”与“场景联合增强”保证扰动在多时序、多视角下保持一致性与有效性,在开放及闭环仿真中显著提升了攻击成功率与隐蔽性.Liu 等人[37]关注任务决策层的语言对抗,发布 EIRAD 多模态数据集并设计有目标/无目标两类策略,以任务分步相似度衡量攻击成效,证明其方法在多项具身任务上以更低迭代代价取得更高成功率.Chen 等人[38]则从物理可部署角度出发,联合优化纹理与透明度生成对抗贴纸,使导航机器人在多视角条件下持续失效.这三类实证结果共同揭示了具身智能在“感知—决策—执行”全链路中的对抗脆弱性,为后续构建鲁棒训练、异常检测与运行时防御提供了可验证的威胁模型与评测基准.

3.6 幻觉放大攻击

幻觉放大攻击指攻击者通过精心构造的输入,引导多模态大型语言模型(Multimodal Large Language Model,MLLM)在生成过程中输出与真实图像内容不符的虚假叙述,并将这些谬误进一步扩展为对象、属性乃至关系层面的系统性错误.与依赖提示词操控、语义误导或物理干扰的传统方法不同,该攻击直接作用于模型内部注意力分布,利用所谓“注意力陷阱”,即模型在几乎无语义载荷的标点符号或连接词上异常集中的注意力,来激活和放大 MLLM 的固有幻觉倾向.当生成流程被迫围绕这些陷阱展开时,模型极易产生与输入图像无关甚至相矛盾的描述,从而在图像描述、视觉问答和任务推理等关键环节削弱具身智能系统的可靠性.

Wang 等人[39]首次系统揭示了注意力陷阱与幻觉生成之间的结构性关联,并提出一种无需预定义目标或模板、仅通过优化中层注意力图和隐藏状态即可诱发高强度幻觉的方法.该两阶段扰动策略在保证输出流畅度的同时,将对象、属性与关系层面的幻觉发生率显著提升.InstructBLIP、MiniGPT-4、LLaVA-1.5 等主流开源模型中多达 75% 的生成内容与原始图像无关;闭源系统(如 GPT-4o、Gemini 1.5 Flash)亦呈现强迁移性.更为严峻的是,该方法能够绕过 OPERA[40]、Less is More[41]、LURE[42]等现有幻觉缓解机制,在多种评测任务中持续保持高效.研究结果不仅验证了“注意力陷阱是幻觉生成核心触发因子”这一假设,也表明即便在无外部恶意指令的场景下,具身系统仍可能因模型内生缺陷而暴露于高危风险,为未来的安全防御提出了全新的挑战.

3.7 跨纬度攻击

跨维度攻击是指攻击者通过协同利用外源性风险与内源性脆弱性而发起的复合型攻击.其中,非侵入式物理对抗攻击是该类攻击的典型代表,其融合了对抗样本生成与传感器欺骗攻击的技术特性[43].此类攻击在外部物理世界中制造扰动并干扰 EAI 系统感知输入,并利用模型内部高维点积的先天缺陷进一步放大扰动效果,最终误导决策输出.物理对抗攻击与在像素层面直接添加扰动的数字对抗攻击不同,物理对抗攻击通过改变实体环境实现干扰,在现实环境中展现出较高的实施可行性和潜在破坏能力[44].根据实施方式的不同,物理对抗攻击可分为侵入式与非侵入式两类.侵入式攻击通过添加实体标记实施干扰,例如在交通标志上粘贴醒目贴纸,此类攻击易于被识别且易受环境或人为因素影响.非侵入式攻击则利用激光、电磁波等物理信号,以人眼难以感知的方式直接干扰传感器数据采集,因而具备更高的隐蔽性.

相关研究已展示了此类攻击的具体实现机制.Sun 等人[45]开发了一种基于激光注入的物理对抗攻击方法,通过训练强化学习模型优化激光束特征,成功误导具身智能体的视觉识别系统,如将限速 90 标志误判为停车标志.该研究在 CARLA 仿真平台的 3 种不同环境中进行了验证,覆盖 4 种识别对象和多种网络架构(如 ResNet、DenseNet、GoogLeNet 等)进行测试,攻击成功率 ASR 超过 60%.Kim 团队[46]提出了一种光学对抗攻击技术,通过在成像系统的傅里叶平面使用空间光调制器对光波相位进行梯度优化,物理层面改变成像传感器的光场分布.该方法可使基于受影响图像的深度学习分类器(如 ResNet、VGG 和 MobileNet)产生误判,同时保持攻击前后图像的视觉一致性.在 ImageNet 数据集和真实光学系统上的实验表明,该攻击能显著降低模型准确率,同时保持较高的图像质量指标.Liu 等人[47]则探索了基于电磁干扰的物理对抗攻击,利用相机传感器对电磁信号的敏感性,注入人眼不可见的精心设计电磁扰动.该研究在 ResNet、YOLO 等多种模型架构上实现了不同类型的攻击,并在 5 种真实摄像头中验证了攻击效果,其中攻击成功率 ASE 在 90% 以上.

跨维度攻击的成功实施,本质上源于从物理环境到数字决策的信息链中存在的传导漏洞. 通过对激光注入、光学干扰及电磁攻击等典型案例[45-47]的分析,可以清晰地勾勒出此类攻击的完整路径:攻击始于物理信号的精准注入,经由传感器端的信号耦合与转换,突破模型的决策边界,最终导致系统出现严重误判. 其中影响攻击有效性的关键因素主要包含 3 个方面:1)扰动信号的物理特性,具体表现为激光参数与电磁频谱的精确控制能力;2)环境与传感器特性对信号传输的调制作用;3)模型决策边界本身存在的结构性脆弱点. 这些因素共同决定了攻击的隐蔽性、鲁棒性与成功率. 针对此类威胁,需构建覆盖感知、模型与系统全链路的协同防御体系. 在感知层,应结合多传感器交叉验证机制,并针对特定物理干扰设计硬件防护措施;在模型层,可基于已识别的攻击模式开展针对性对抗训练,同时建立特征异常监测机制以识别潜在干扰;在系统层,需部署实时决策逻辑监控模块,实现对异常行为的及时阻断与控制策略切换. 通过打破传统防御组件间的隔离状态,构建基于攻击机理深度理解的跨层安全架构,逐步形成具备内在韧性的系统安全能力.

4　安全防御机制

具身智能体在感知、决策与执行的完整链路中,面临多种安全威胁,例如后门触发、传感器欺骗、对抗扰动、越狱提示以及模型幻觉等. 然而,面向具体智能体的安全防护研究仍处于初期阶段,这是因为物理环境本身的高度复杂性与模型内在的不确定性对 EAI 任务执行的精确性与可靠性构成严峻挑战. 所以当前工作更多聚焦于安全性(safety)问题即防止意外事故或自然风险导致的安全状况,而非广义上侧重人为恶意行为和外部威胁的安全防护(security). 举例来说,幻觉放大攻击的核心本质是因为模型内部存在极大的不确定性,因此当下的工作偏向于先降低EAI决策过程不确定性对任务执行的影响,再去考虑是否如何检测并缓解这种不确定性被恶意利用的情形. 近年来已出现将安全防御机制嵌入具身智能体的初步尝试,推动其走向实际部署. 随着该领域的发展,随着相关研究的推进,防御体系逐步从形式化约束、可达性验证、多模态反馈以及基准评测等 4 个主要方向展开,致力于构建系统化的防御体系. 表 1 概括了代表性工作及其覆盖的威胁面,以下各节将详细阐述其核心思想及最新进展.

表 1　具身智能决策安全防御技术总览表

Table 1　Overview of security and defense technologies for embodied intelligent decision making

类别	代表方法	机制	优势
形式化约束	Safe Planner[48] Safety Chip[49]	使用 PDDL 或 LTL 形式定义任务约束,规划过程中动态检查或生成合法策略.	可验证性强,规则明确,配合已有标准.
可达性验证	Reachability Analysis[50]	预测系统未来状态构成可达集,判断动作是否将引发危险.	不依赖模型内部结构,适合底层控制层.
多模态反馈	TrustNavGPT[52] SafeEmbodAI[53]	声音/图像等模态结合提示词分析,引入置信判断与行为撤销.	可实时检测自然语言越狱与误导.
基准评测	SafeAgentBench[54] AGENTSAFE[55]	在任务与动作两层设置规则/监控,对多步计划进行评估与纠错.	可用于训练数据审查与执行前后验证.

4.1　基于形式化约束机制的任务规划安全增强

当大型语言模型生成具身任务计划时,往往缺乏对动态环境与安全规则的深层理解,容易输出语义正确却物理危险的操作序列. 因为缺乏对物理规则的深层理解,所以更易受到模型幻觉的影响. 为此,研究者首先在高层规划环节引入显式约束,使安全意识伴随推理全过程. Li 等人[48]建立了“任务—技能”二级框架,在高层使用 EAI 生成候选方案的同时,由低层的安全预测网络对每条候选技能进行评分,规划器再依据评分优先选择低风险路径,并通过规划领域定义语言(Planning Domain Definition Language,PDDL)确保约束的结构化表达. 与之互补,Yang 等人[49]将线性时序逻辑(Linear Temporal Language,LTL)直接嵌入语言提示的解析流程,利用自动机将自然语言映射为可验证的形式表达,并在推理过程中实时监控约束可达性,一旦检测到违例立即触发规划重写. 通过引入显式约束规则并强化对动态环境与安全规则的理解,可以有效降低由于环境不对齐攻击带来的潜在风险. 在任务规划过程中,模型能够更加注重物理可行性和安全性,从而提升在实际执行中的可靠性和稳定性.

尽管形式化约束可以在一定程度上缓解幻觉带来的影响[49]. 但幻觉放大攻击依然是一个极为棘手且重要的挑战. 这是因为形式化约束通常依赖于对模型状态空间和控制流程的精确建模,而幻觉放大攻击本质上通过操控模型内部的注意力机制或跨模态交互引发不可预测的错误行为. 此类攻击依赖于模型内生的认知偏差,现有的防御方法往往无法单独有效应对这一问题. 为了更有效地抵御幻觉放大攻击,本文认为需要在数据、训练和推理多个阶段同时进行针对性的缓解和限制. 首先,通过提高数据质量减少模型内部已有的错误. 其次,改进预训练策略增强模型的上下文理解能力,探索新的训练范式提升模型对复杂任务的适应能力并研究净化对抗扰动的防御措施. 最后,在推理阶段,除了结合形式化约束进行事实验证外,还需要开发有效的防御系统专门针对幻觉进行抑制,确保模型生成的任务计划语义正确并符合实际可行性和安全性要求.

4.2　基于可达性验证的控制轨迹修正

即便高层计划满足安全约束,EAI 生成的低层动作仍需与实时环境和动力学模型对齐,否则可能在执行环节引发碰撞或失稳. 传统形式化约束依赖精确建模,难以覆盖真实机器人动力学的复杂性,因而近期工作转向数据驱动的可达性验证. Hafez 等人[50]借助历史交互数据构造状态动作可达集,在运行时通过外推预测下一步动作的安全边界;若预测轨迹与障碍区域重叠,系统立即裁剪动作或调用应急策略,形成“反馈—可达—剪枝”的闭环. 与此同时,可学习屏障函数将控制安全边界参数化为可训练网络,通过强化学习在真实轨迹中不断校准,使屏障约束自适应 EAI 输出的长尾分布. 两种方法从数据而非解析模型出发,为复杂动力学环境提供了实时、

可扩展的安全验证手段,与高层约束形成上下呼应.

可行性分析整合了控制学理论,因此可以计算未来状态来确保模型安全,在应对复杂动力学环境中的攻击挑战时,该方法具有极强的前景[51]. 通过对当前场景构造动态模型,可达性验证能够实时监控和修正轨迹来降低攻击者注入的扰动的影响. 举例来说,当攻击者使用对抗攻击、传感器欺骗等手段注入扰动后,模型可能会忽视前方的障碍物,并发出指令操控智能体撞击到障碍物上,但是可行性分析会预估这种可能性,通过外推预测未来动作的安全边界并与实际障碍区域进行对比,系统在预测到潜在的危险时立即调整动作,从而减少由扰动引发的偏差. 即使传感器数据被篡改,系统仍然可以通过历史数据和当前动作的可达性分析,预测并修正 EAI 的执行轨迹,确保控制行为与实际环境一致. 通过动态调整控制边界并持续校准,系统能应对传感器数据的异常,减少传感器欺骗攻击对决策过程的干扰.

4.3 基于多模态反馈的风险感知与拒绝机制

形式化约束和可达性验证这两种高层与低层的安全增强机制虽能削减大部分显性风险,但自然语言本身的模糊不确定性或者潜在的恶意仍可能诱使系统越狱或触发幻觉. 为增强交互安全,研究者进一步在感知与交互环节引入多模态不确定性建模与拒绝机制. Sun 等人[52]在移动机器人平台上同时分析文本指令的语义模糊度与语音信号的犹豫特征(例如停顿和音高变化),当综合不确定性超过阈值时,机器人选择降低速度或请求人机澄清,以此避免在含糊场景中贸然执行. Zhang 等人[53]通过 Secure Prompting 将结构化规则库与状态记忆块嵌入多模态交互循环,在生成计划前先检查指令是否触犯禁止区域或与历史状态冲突;若发现潜在风险,系统会主动拒绝或改写提示,从源头拦截越狱尝试. 上述工作表明,风险感知与拒绝机制不仅能补强前两节的结构性保障,还可在复杂交互场景中维持人机协作的鲁棒性与可接受性.

在应对多种攻击方式时,多模态反馈机制对对抗攻击与越狱攻击具备一定的缓解作用. 首先,对于对抗攻击,多模态反馈通过结合语音、文本和视觉信息的不确定性(如文本语义模糊度、语音中的犹豫特征、视觉置信度等)进行综合评估,当系统检测到潜在的异常信号时,会通过降低执行速度、请求进一步的澄清或暂缓执行等策略来控制风险,避免对抗扰动对决策产生过大影响. 其次,对于越狱攻击,通过引入安全机制,系统能够在生成任务计划前对输入指令进行多模态检查和历史状态验证. 当指令与历史状态或安全规则存在冲突时,系统将主动拒绝或重写指令,从而有效地阻止潜在的恶意操作. 这些多模态反馈机制通过在交互过程中实时识别和处理潜在的安全威胁,加强了系统的安全性,并补充了传统的结构性约束和可达性分析方法.

然而,单纯依赖多模态反馈机制难以有效抵御幻觉放大攻击. 这类攻击本质上源于模型内部的认知偏差,而多模态反馈主要基于外部感知信息进行风险评估,无法直接介入模型内部的生成逻辑. 具体而言,幻觉放大攻击作用于模型的内容生成与推理阶段,使 EAI 在外部输入无明显异常的情况下,仍可能因注意力机制偏移或语义推理偏差而产生错误的任务规划. 多模态反馈机制通常仅在最终决策输出环节实施干预,缺乏对生成过程中间状态(如中间层表示、注意力权重等)的细粒度监控与控制,因而难以及时察觉并抑制生成过程中逐步形成的认知偏差. 此外,幻觉放大攻击所生成的错误输出在表面上常保持多模态间的一致性,使得基于跨模态一致性的检测机制也难以识别其内在的推理缺陷或语义矛盾.

4.4 安全基准体系与行为评测框架

当前大多数具身智能研究侧重任务成功率,忽视了系统在执行中可能引发的安全风险. 最近两年,学界陆续发布了面向具身安全的公开基准,为不同策略提供了可复现、可对比的测试平台. 这些基准工具不仅推动具身智能安全研究从定性走向定量,也为未来系统性对比不同防御策略、训练鲁棒模型提供了基础设施与标准数据.

Yin 等人[54]构建了 SafeAgentEnv,这是一个基于 3D 仿真环境 AI2-THOR 的交互式 EAI 任务场景,旨在涵盖电机故障、滑倒、爆炸等潜在的安全风险事件. 该基准通过提供细粒度的语义标签和执行日志,便于验证部署大语言模型的 EAI 在长时序任务中的表现. 在该环境中执行复杂任务时,研究发现现有 EAI 系统的安全意识较为薄弱. 特别是在处理具有高危险性的任务时,使用 GPT-4 驱动的 ReAct 模型的任务拒绝率仅为 10%.

Ying 等人[55]提出了 AGENTSAFE,这是一个用于评估视觉语言模型(VLM)驱动的 EAI 在执行危险指令时安全性的方法. AGENTSAFE 基于 AI2-THOR 构建了一个交互式场景,包含了 1350 个危险任务和 8100 个危险指令的数据集,用于模拟 EAI 在感知、规划和执行阶段可能面临的安全风险. 该基准测试评估了多种主流 VLM(包括 GPT、Claude、Gemini 等)驱动的 EAI. 尽管这些智能体在执行一些基础安全任务时表现良好,但在面对越狱攻击和危险指令时,许多智能体未能有效拒绝不安全的任务. 例如,GPT-4 驱动的 ReAct 模型在面对可能损坏环境或危害自身安全的任务时,其拒绝率低于 10%.

现有的安全基准和评估方法在对具身智能体安全性的定量评估中发挥了重要作用. 尽管 LLM 驱动的 EAI 在一些任务中表现出较好的执行能力,但在处理复杂和高风险的任务时,其安全意识仍然存在不足,可靠性亟待进一步提升.

5 问题与展望

具身智能的安全研究虽然已从零散讨论走向体系化探索,但与大规模实际部署的安全需求相比仍然存在显著缺口. 通过回顾现阶段的工作,可以发现若干制约进一步发展的核心瓶颈. 展望未来,解决这些问题不仅需要算法与硬件的协同创新,也离不开多学科融合以及产业与监管的共同推进.

5.1 亟待突破的核心问题

5.1.1 语义与物理的安全脱节问题严重

语义与物理的安全验证存在显著脱节. 现有端到端大模型所主导的 EAI 主要关注自然语言指令的合规性与逻辑合理性,对生成策略在物理环境中可行性与安全性进行验证的手段尚不完善,导致语义层面正确的决策可能在物理执行阶段产生不可行或危险的行为序列. 具体表现为任务规划阶段未能充分考虑动态环境约束、机器人本体物理限制以及执行过程中的不确定性因素,引发碰撞、失稳等物理安全隐患. 在

高风险场景中,此类问题可能进一步导致严重安全事故,比如在自动驾驶场景中无人车违反交通法规、引发车祸造成人员伤亡等. 该问题的本质在于尚未建立有效的语义逻辑与物理规则对齐机制,难以全面覆盖实际应用中的潜在异常情况.

5.1.2　防御策略覆盖维度有限,抗越界能力弱

当前安全机制多针对特定环节或单一攻击类型进行设计,例如输入过滤、输出校正或局部轨迹优化等,缺乏跨层级、跨模块的系统性防护能力. 这种分散的防御架构使 EAI 在面对复杂的跨维度攻击时表现出明显脆弱性,特别是难以应对涉及感知、决策与执行链路的组合式攻击. 各防御组件之间缺乏有效的信息共享与联动机制,无法形成统一的威胁感知与响应体系,导致攻击者能够通过多层级渗透绕过局部防护. 这一系统性缺陷反映了当前研究尚未建立覆盖全生命周期与全业务流程的整体安全框架.

5.1.3　安全验证与端到端模型耦合松散

当前的安全验证方法往往作为独立于核心决策流程的后置模块存在,导致验证环节与模型推理过程形成明显的断层和隔阂. 这种架构缺陷具体表现为:验证模块难以及时获取决策过程中的中间状态和语义信息,而决策模型也无法充分利用验证结果进行在线调整和优化. 现有方法在验证覆盖率与系统性能之间的平衡尚未得到有效解决,难以在不影响任务性能的前提下,将非结构化的安全需求转化为模型内部可理解的约束知识. 这种松散的集成架构大大降低了安全验证的时效性与有效性.

5.1.4　开放环境下的风险感知缺乏

当前风险识别机制主要针对预设场景和已知威胁进行优化,缺乏对动态开放环境中新型威胁的主动发现能力. 具体局限体现在 3 个方面:系统难以有效识别非结构化环境中的突发安全威胁;在复杂多变场景中对潜在风险的检测存在高延迟;现有风险感知模块缺乏对多模态信息的深度融合能力,无法基于环境动态、任务上下文和系统状态等多维度数据构建完整的风险态势认知. 这些能力缺陷的根源在于当前评估、仿真和训练过程依赖于有限场景下的静态假设,未能建立适应开放环境动态特性的先验风险感知框架.

5.2　未来研究趋势与重点方向

基于当前具身智能安全领域面临的核心问题,未来研究应重点围绕语义物理对齐、跨层协同、端到端可验证框架和先验风险感知 4 个关键方向展开系统性突破,并建立清晰的技术发展路径.

在语义物理对齐方面,需构建连接语义推理与物理验证的统一框架. 该方向面临语义符号与物理规则间缺乏关联、对齐偏差导致模型幻觉等核心挑战,属于中长期攻坚方向. 建议采取分阶段实施路径:近期重点建立基础语义-物理关联模型,在受限环境中验证关键技术;中期拓展到复杂动态场景,实现语义规划与物理执行的闭环验证;长期目标是形成成熟的理论体系和技术标准,为具身智能的可靠部署提供基础支撑. 在跨层协同防御方面,需重塑当前碎片化的防御架构. 重点研究方向包括建立跨层威胁信息共享机制、设计协同响应策略以及构建统一的防御框架. 实施路径可从定义标准化交互接口起步,逐步建立跨层协同的防御原型,最终形成自适应智能防御体系. 端到端可验证框架是确保系统可靠性的关键保障. 该方向需要突破形式化验证技术与深度学习模型的融合瓶颈,解决验证过程的可扩展性和实时性问题. 技术发展应遵循从离线验证到在线监控、从组件级验证到系统级保证的演进路径,构建统一的验证理论和工具链. 在先验风险感知方面,需发展面向开放环境的动态风险识别能力. 重点包括构建多模态风险感知架构、开发在线学习机制以及建立风险预测模型. 建议采取理论创新-算法突破-系统实现的递进式发展路径,通过持续的知识积累和技术迭代,逐步提升 EAI 对未知风险的应对能力.

需要强调的是,这 4 个方向并非孤立发展,而是构成了相互支撑的有机整体. 语义物理对齐为系统提供基础安全保障,跨层协同防御确保防护体系的整体效能,端到端验证框架建立可信性基石,先验风险感知则赋予系统动态适应能力. 未来研究应当注重各方向之间的协同创新,通过建立层次化、阶段化的技术发展路线图,系统推进具身智能安全能力的全面提升.

6　结　论

具身智能作为通往通用人工智能的重要途径,其决策安全已成为制约现实部署与产业化落地的关键瓶颈. 本文围绕大模型驱动的具身智能决策逻辑安全这一核心议题,对近几年公开文献进行了系统梳理,首先梳理了具身智能的基本概念与发展脉络,随后提出面向输入干扰、模型脆弱和指令操控的决策风险分析框架,并从外部攻击(后门、传感器欺骗、越狱)与内部缺陷(环境不对齐、对抗扰动、幻觉放大)两条主线详细剖析了典型威胁机理. 在此基础上,本文总结了基于形式化约束、可达性验证、多模态反馈与基准评测的 4 类防御思路,分析其优势与局限.

综合现有研究,虽然大语言模型与多模态大模型赋予具身系统更强的任务泛化与语义推理能力,但模型的概率性推理、黑箱性结构和开放式交互又拓宽了攻击面. 当前防御方法仍主要分散在感知过滤、输出修补或局部验证等局部手段,尚未形成贯穿“感知—决策—执行”的安全闭环;同时因为缺乏对策略可执行性、动力学可达性与伦理合规性的统一评估,实验结果难以直接转化为真实场景可用保障. 展望未来,具身安全研究可以重点围绕语义物理对齐、跨层协同、端到端可验证框架和先验风险感知 4 个关键方向展开系统性突破. 将 EAI 决策安全机制内嵌到模型规划逻辑与执行链路,并辅以系统化评测与合规监管,具身智能有望在服务、工业、医疗等高价值场景中实现安全、可靠的广泛应用.

References:

[1] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[2] WANG W S, TAN N, HUANG K, et al. Embodied intelligence systems based on large models: a survey[J]. Acta Automatica Sinica, 2025, 51(1): 1-19.

[3] Wake N, Kanehira A, Sasabuchi K, et al. GPT-4V(ision) for robotics: multimodal task planning from human demonstration[J]. IEEE Robotics and Automation Letters, 2024, 9(11): 10567-10574.

[4] Lu D, Sun Y, Zhang Z, et al. InternVL-X: advancing and accelera-

ting internVL series with efficient visual token compression[J]. arXiv preprint arXiv:2503.21307,2025.

[5] Team G,Georgiev P,Lei V I,et al. Gemini 1.5:unlocking multimodal understanding across millions of tokens of context[J]. arXiv preprint arXiv:2403.05530,2024.

[6] Liu Y,Chen W,Bai Y,et al. Aligning cyber space with physical world:a comprehensive survey on embodied AI[J]. arXiv preprint arXiv:2407.06886,2024.

[7] Cangelosi A,Bongard J,Fischer M H,et al. Embodied intelligence[M]. Springer Handbook of Computational Intelligence,Springer Nature,2015:697-714.

[8] Xing W,Li M,Li M,et al. Towards robust and secure embodied AI:a survey on vulnerabilities and attacks[J]. arXiv preprint arXiv:2502.13175,2025.

[9] Turing A M. Computing machinery and intelligence[M]. Springer Netherlands,2009.

[10] Liu H,Guo D,Cangelosi A. Embodied intelligence:a synergy of morphology,action,perception and learning[J]. ACM Computing Surveys,2025,57(7):1-36.

[11] XU W Y,JI X Y,YAN C,et al. Embodied artificial intelligence security and governance[J]. Bulletin of Chinese Academy of Sciences,2025,40(3):429-439.

[12] Radford A,Kim J W,Hallacy C,et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning,2021:8748-8763.

[13] Ahn M,Brohan A,Brown N,et al. Do as i can,not as i say:grounding language in robotic affordances[J]. arXiv preprint arXiv:2204.01691,2022.

[14] Li J,Li D,Xiong C,et al. Blip:bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning,2022:12888-12900.

[15] Li J,Selvaraju R,Gotmare A,et al. Align before fuse:vision and language representation learning with momentum distillation[C]//Advances in Neural Information Processing Systems,2021:9694-9705.

[16] Ho J,Ermon S. Generative adversarial imitation learning[C]//Advances in Neural Information Processing Systems,2016:4572-4580.

[17] Puig X,Undersander E,Szot A,et al. Habitat 3.0:a co-habitat for humans,avatars and robots[J]. arXiv preprint arXiv:2310.13724,2023.

[18] Brohan A,Brown N,Carbajal J,et al. Rt-1:robotics transformer for real-world control at scale[J]. arXiv preprint arXiv:2212.06817,2022.

[19] Schulman J,Wolski F,Dhariwal P,et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347,2017.

[20] Roderick M,MacGlashan J,Tellex S. Implementing the deep q-network[J]. arXiv preprint arXiv:1711.07478,2017.

[21] Haarnoja T,Zhou A,Hartikainen K,et al. Soft actor-critic algorithms and applications[J]. arXiv preprint arXiv:1812.05905,2018.

[22] Xiang,Zhen,et al. Backdoor chain-of-thought prompting for large language models[J]. arXiv preprint arXiv:2401.12242,2024.

[23] Wang X,Pan H,Zhang H,et al. Trojanrobot:backdoor attacks against robotic manipulation in the physical world[J]. arXiv eprints,2024:arXiv:2411.11683.

[24] Jiao R,Xie S,Yue J,et al. Can we trust embodied agents? exploring backdoor attacks against embodied LLM-based decision-making systems[J]. arXiv preprint arXiv:2405.20774,2024.

[25] Liu A,Zhou Y,Liu X,et al. Compromising LLM driven embodied agents with contextual backdoor attacks[J]. IEEE Transactions on Information Forensics and Security,2025,20:3979-3994,doi:10.1109/TIFS.2025.3555410.

[26] Ji X,Cheng Y,Zhang Y,et al. Poltergeist:acoustic adversarial machine learning against cameras and computer vision[C]//IEEE Symposium on Security and Privacy(SP),2021:160-175.

[27] Jin Z,Ji X,Cheng Y,et al. Pla-lidar:physical laser attacks against lidar-based 3d object detection in autonomous vehicle[C]//IEEE Symposium on Security and Privacy(SP),2023:1822-1839.

[28] Jang J H,Cho M,Kim J,et al. Paralyzing droncs via EMI signal injection on sensory communication channels[C]//Network and Distributed System Security Symposium(NDSS),2023,doi:10.14722/ndss.2023.24616.

[29] Robey A,Ravichandran Z,Kumar V,et al. Jailbreaking LLM-controlled robots[J]. arXiv preprint arXiv:2410.13691,2024.

[30] Zhang H,Zhu C,Wang X,et al. Badrobot:jailbreaking LLM-based embodied ai in the physical world[J]. arXiv preprint arXiv:2407.20242,2024.

[31] Lu X,Huang Z,Li X,et al. POEX:policy executable embodied AI jailbreak attacks[J]. arXiv preprint arXiv:2412.16633,2024.

[32] Robey A,Ravichandran Z,Kumar V,et al. Jailbreaking LLM-controlled robots[J]. arXiv preprint arXiv:2410.13691,2024.

[33] Tan W,Zhang W,Liu S,et al. True knowledge comes from practice:aligning LLMs with embodied environments via reinforcement learning[J]. arXiv preprint arXiv:2401.14151,2024.

[34] Du Y,Watkins O,Wang Z,et al. Guiding pretraining in reinforcement learning with large language models[C]//International Conference on Machine Learning,2023:8657-8677.

[35] Zou A,Wang Z,Carlini N,et al. Universal and transferable adversarial attacks on aligned language models[J]. arXiv preprint arXiv:2307.15043,2023.

[36] Zhang T,Wang L,Zhang X,et al. Visual adversarial attack on vision-language models for autonomous driving[J]. arXiv preprint arXiv:2411.18275,2024.

[37] Liu S,Chen J,Ruan S,et al. Exploring the robustness of decision-level through adversarial attacks on LLM-based embodied models[C]//Proceedings of the 32nd ACM International Conference on Multimedia,2024:8120-8128.

[38] Chen M,Tu J,Qi C,et al. Towards physically-realizable adversarial attacks in embodied vision navigation[J]. arXiv preprint arXiv:2409.10071,2024.

[39] Wang Y,Zhang M,Sun J,et al. Mirage in the eyes:hallucination attack on multi-modal large language models with only attention sink[J]. arXiv preprint arXiv:2501.15269,2025.

[40] Huang Q,Dong X,Zhang P,et al. Opera:alleviating hallucination in multi-modal large language models via over-trust penalty and

retrospection-allocation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2024:13418-13427.

[41] Yue Z,Zhang L,Jin Q. Less is more:mitigating multimodal hallucination from an eos decision perspective[J]. arXiv preprint arXiv:2402.14545,2024.

[42] Zhou Y,Cui C,Yoon J,et al. Analyzing and mitigating object hallucination in large vision-language models[J]. arXiv preprint arXiv:2310.00754,2023.

[43] Fang J,Jiang Y,Jiang C,et al. State-of-the-art optical-based physical adversarial attacks for deep learning computer vision systems[J]. Expert Systems with Applications, 2024, 250: 123761-123771,doi:10.1016/j.eswa.2024.123761.

[44] Wei H,Tang H,Jia X,et al. Physical adversarial attack meets computer vision:a decade survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2024,46(12):9797-9817.

[45] Sun Y,Huang Y,Wei X. Embodied laser attack:leveraging scene priors to achieve agent-based robust non-contact attacks[C]//Proceedings of the 32nd ACM International Conference on Multimedia,2024:5902-5910.

[46] Kim K,Kim J,Song S,et al. Engineering pupil function for optical adversarial attacks[J]. Optics Express,2022,30(5):6500-6518.

[47] Liu Z,Lin F,Ba Z,et al. MagShadow:physical adversarial example attacks via electromagnetic injection[J]. IEEE Transactions on Dependable and Secure Computing,2025,22(4):3307-3323.

[48] Li S,Liu F,Cui L,et al. Safe planner:empowering safety awareness in large pre-trained models for robot task planning[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2025:14619-14627.

[49] Yang Z,Raman S S,Shah A,et al. Plug in the safety chip:enforcing constraints for LLM-driven robot agents[C]//IEEE International Conference on Robotics and Automation (ICRA),2024:14435-14442.

[50] Hafez A,Akhormeh A N,Hegazy A,et al. Safe LLM-controlled robots with formal guarantees via reachability analysis[J]. arXiv preprint arXiv:2503.03911,2025.

[51] Tan X,Liu B,Bao Y,et al. Towards safe and tustworthy embodied AI:foundations,status,and prospects[EB/OL]. https://openreview.net/pdf? id=Eu6Yt21Alv,2025-09-12.

[52] Sun X,Zhang Y,Tang X,et al. TrustNavGPT:modeling uncertainty to improve trustworthiness of audio-guided LLM-based robot navigation[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS),2024:8794-8801.

[53] Zhang W,Kong X,Braunl T,et al. Safeembodai:a safety framework for mobile robots in embodied ai systems[J]. arXiv preprint arXiv:2409.01630,2024.

[54] Yin S,Pang X,Ding Y,et al. SafeAgentBench:a benchmark for safe task planning of embodied LLM agents[J]. arXiv preprint arXiv:2412.13178,2024.

[55] Liu A,Ying Z,Wang L,et al. AGENTSAFE:benchmarking the safety of embodied agents on hazardous instructions[J]. arXiv preprint arXiv:2506.14697,2025.

附中文参考文献:

[2] 王文晟,谭　宁,黄　凯,等.基于大模型的具身智能系统综述[J].自动化学报,2025,51(1):1-19.

[11] 徐文渊,冀晓宇,闫　琛,等.具身智能安全治理[J].中国科学院院刊,2025,40(3):429-439.