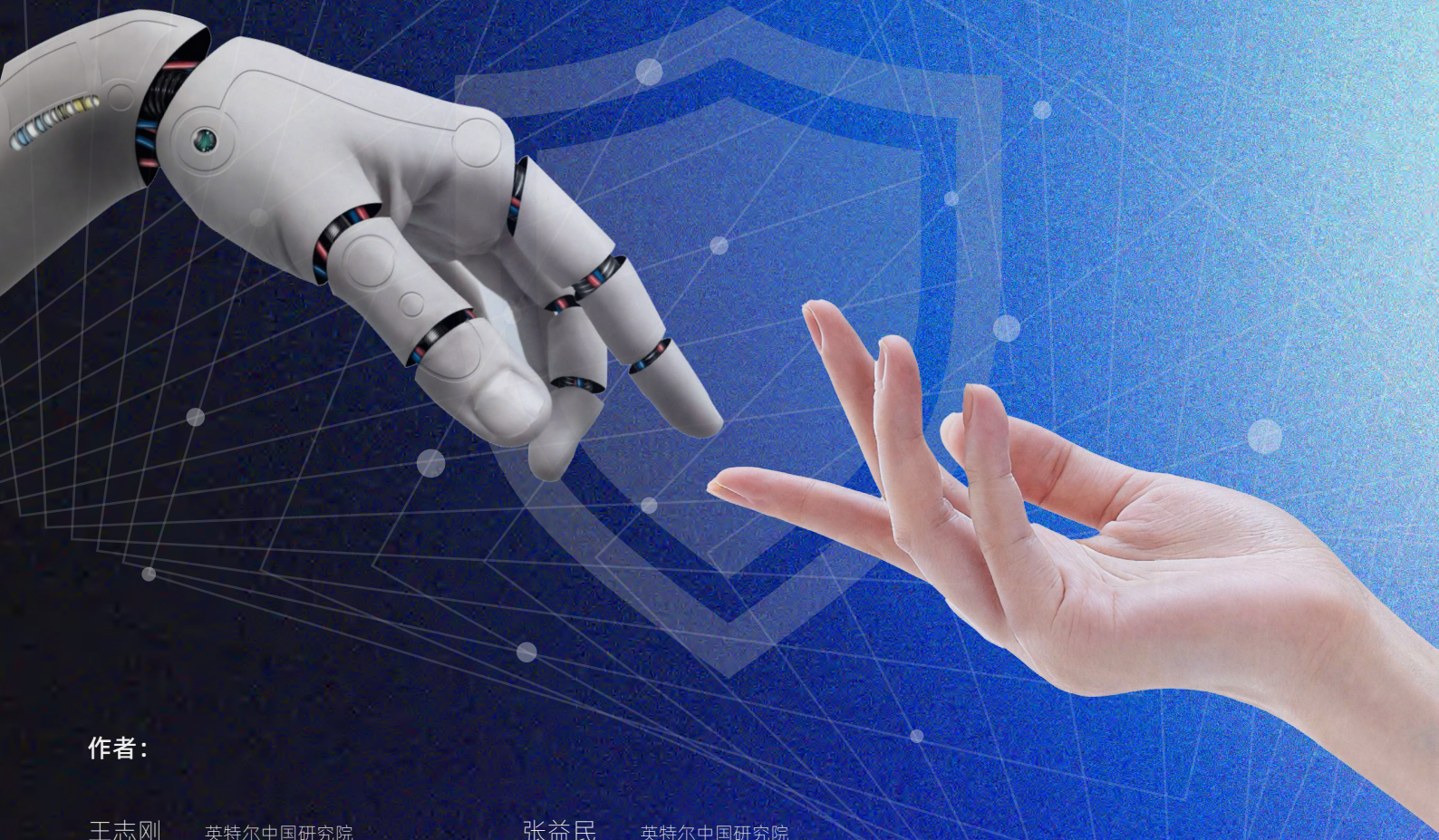
具身智能机器人安全子系统白皮书



作者：

王志刚　英特尔中国研究院

赵明国　清华大学

张强　国地共建具身智能机器人创新中心

陈勇全　香港中文大学（深圳）

宋继强　英特尔中国研究院

张益民　英特尔中国研究院

李淼　武汉大学机器人学院

庞建新　深圳市优必选科技股份有限公司

吴冶　南京英麒智能科技有限公司

杨洪　英特尔亚太研发有限公司

2025 年 12 月

目　录

1. 前言

在人工智能技术浪潮的推动下，具身智能机器人正以前所未有的速度融入我们的生产与生活。具身智能机器人，指的是能够集成感知、认知与执行能力，在真实环境中自主决策并完成物理交互任务的智能体，是本体（机器人实体）、环境、智能能力三者深度耦合形成的“三位一体”协同系统。它既能被动感知环境，又可以主动影响环境，在与环境的动态耦合中，实现“感知->认知->决策->行动->反馈”的闭环感知与行为学习。相较于传统工业机器人，这类机器人具有更强的环境适应性、更高的灵活性，以及更复杂的人机交互能力。它们不仅涵盖人形机器人，还包括仿生机器人、移动服务机器人等多种形态，广泛应用于制造、医疗、家庭服务、公共场所等领域。

可以预见，在不远的未来，具身智能机器人将在工厂中承担复杂的搬运与装配任务，成为生产线上的得力助手；在家庭中协助完成家务、照料老人，显著提升生活便利性与福祉水平；在公共场所中提供导引、咨询等服务，提高公共服务的效率与质量。

近年来，人形机器人作为具身智能机器人的一个典型形态，因其类人的外观与运动方式，在协作任务与自然交互方面展现出独特优势。它们能够更自然地与人类沟通、协作，适应人类工作和生活场景，已成为具身智能研究和产业发展的热点。

然而，随着不久的将来机器人包括人形机器人逐渐成为人类身边的“常客”，其安全性愈发成为悬在头顶的“达摩克利斯之剑”。想象一下，一个在家庭中帮助看护老人的人形机器人，如果失去控制，可能会撞倒老人；或者在做饭时，因恶意操控而拿着菜刀对人进行攻击。如何才能像阿西莫夫机器人三定律所设想的那样，让机器人不伤害人和环境呢？实现可靠的安全方案已成为当务之急。

本文将在[1]的基础上，以人形机器人作为一个重点去探讨具身智能机器人的安全，而所有的讨论都适用于具身智能机器人(包括人形机器人和非人形的如移动服务机器人等)，下面的讨论中对两者不再做具体区分。

具身智能机器人的安全性是一个多维度的问题，主要涉及以下四个核心维度。

- 信息安全，是机器人的“数字防线”。在这个万物互联的时代，机器人通过网络与外部世界时刻交换信息，涉及大量敏感数据。一旦遭遇黑客攻击，这些信息可能被窃取或篡改，导致个人隐私泄露、商业机密被盗，甚至可能被恶意利用来实施犯罪行为，给个人、企业乃至社会带来巨大损失。更为重要的是，信息安全问题也会蔓延至系统软件层，间接影响物理执行层面的安全。

- 功能安全，是守护机器人系统的“可靠根基”。功能安全强调的是在系统发生预期外事件（如传感器故障、控制算法失效、电源异常等）时，仍能通过设计保障系统保持在“可控、可预测”的安全状态。其核心理念是“故障安全（fail-safe）”——即使部分组件失效，机器人仍不会造成伤害或灾难性后果。例如，紧急制停机制、任务降级执行、系统自检测与冗余设计，都是实现功能安全的重要手段。功能安全不仅涵盖电子电气系统，也包括系统控制逻辑、软件行为约束及与环境交互的稳定性。

- 物理安全，是人机共处的“生命红线”。机器人作为物理实体，与人类共享空间时，其体积、重量、关节速度、惯性等物理属性可能带来直接威胁。例如，机械臂制动失效可能导致撞击事故，陪护机器人力量控制不当可能造成老人跌倒。物理安全不仅仅是硬件结构的安全（如材料、结构、绝缘），还包括接触过程中的动态安全性。

- 交互安全[2]，是人机交互的“情境护栏”。随着具身机器人在家庭、医院、公共场所中普及，人与机器的互动越来越频繁，交互安全成为新的关注重点。它涵盖多模态感知误判、语言/手势误识、情绪响应不当、决策逻辑错误等问题。例如，机器人误判儿童动作为威胁而采取错误防御动作，或在语言指令歧义下做出危险操作。此外，决策系统本身的鲁棒性和抗干扰性，也是交互安全的重要保障。

在上述四大维度中，功能安全是系统可靠运行的基石，信息安全是保障其对外连接的防线，物理安全直接关系人身安危，而交互安全则是准确完成任务，实现和谐共处的关键纽带。四者交织，共同构成具身智能机器人安全系统的整体防护体系。

本文将聚焦于功能安全与物理安全，结合交互安全的核心内容[1]，从系统架构层面提出一个安全子系统的设计框架。我们将探讨具身智能机器人中安全子系统与主控系统的接口机制，解析核心模块组成与实现逻辑，并展望具身智能机器人安全技术的未来发展方向，目标是在不损害运动性能的前提下，实现在复杂场景下的人机“零伤害”共处。

[1]信息安全主要涉及通信加密、数据访问控制、网络攻击防护等内容，其技术路线与具身智能机器人本体系统的感知–控制–执行链路关联度相对较低，且已有较为完备的研究基础。本文重点聚焦功能安全与物理安全在具身智能架构中的集成挑战，故不做展开。

2. 具身智能机器人安全的挑战

2.1 物理接触风险

当具身智能机器人与人类近距离交互协作时，诸多物理接触风险亟待解决：

- **快速运动风险 ：** 机器人整体或部分关节若进行快速运动，容易引发碰撞事故，或者导致部件飞出，对人或环境造成伤害。

- **碰撞、挤压风险 ：** 机器人在直接接触人体或环境物体时，可能造成损害。例如，在狭窄通道中搬运重物的机器人，若未精准感知周围环境，就可能与墙壁或人发生碰撞，导致人受伤或物品损坏。当机器人跌倒时，其身体或部件可能会碰撞、挤压人体或环境物体，造成伤害。

- **夹伤风险 ：** 人的手指或衣物很容易被机器人手指、关节缝隙或可动部件（如旋转腰部）夹住，造成机械性损伤。比如，机器人的手指在抓取物体时，若未准确判断位置或抓取目标，就可能夹伤人的手指。

- **手持锐器风险 ：** 当机器人手持刀具等尖锐工具，尤其在并非执行需要用刀的任务（如制作食品）时，存在极高危险。若机器人动作失控或路径规划失误，锐器可能会误伤人。

2.2 间接风险

如果安全风险不是机器人本身操作或运动直接物理接触人或环境物体造成的，就可以归类为间接风险。这方面的风险扩展开来应该非常多，我们这里仅仅列出两类常见的例子：

- **环境风险 ：** 人或机器人可能将果皮或液体等杂物丢弃在地上，导致机器人或人滑倒摔倒。

- **操作水火电气不当风险 ：** 机器人倒开水时若操作失误或被碰撞，可能会烫到人身体；机器人打开煤气后忘记关闭，可能引发燃气泄漏等严重安全事故。

2.3 风险产生原因分析

具身智能机器人，特别是人形机器人是一个复杂的软硬件系统，这方面已经有不少深入的调研如[3]。从硬件角度看，机器人本体由各种传感器、伺服器、计算平台、电池等组成。工业机器人的可靠性通常用平均无故障运行时间（Mean Time Between Failures -- MTBF）表示，国际上主流工业机器人产品的MTBF约在 7.5 万小时到 10 万小时之间，换算成每年的故障率在 8.7% 到 11.3% 之间。家用的人形机器人部件，包括机械臂等，由于成本等因素，其故障率可能会更高，虽然可以通过定期替换部件来降低故障率，但故障依然难以完全避免。

从软件层面而言，最新的人形机器人系统一般由操作系统、中间件以及提供感知、规划、操作相关功能的大小脑模型组成。在这个复杂的系统中，每个硬件部件和软件模块都有可能成为系统出现问题的源头。例如，软件遭受黑客攻击，就可能引发故障。更为关键的是，人形机器人的软件基于人工智能，具身智能系统中的人工智能部分由于本身的成熟度和可靠性问题，如大模型的幻觉问题（即生成不符合实际或错误的信息）、模型的漏洞被恶意攻击等，也会产生安全风险。

如上所述，安全风险可能是机器人本体或软件方面多方面故障造成的。具体说会导致物理安全风险的机器人系统的一些常见故障原因如下：

1. 执行器故障（电机）
- 电机过热/卡死：导致关节突然锁死或失控，可能引发剧烈碰撞或摔倒。
- 扭矩异常：输出力矩超出预期，损坏被操作物体或伤害附近人员。

2. 传感器故障（发生于如力觉、视觉和IMU等传感器）
- 力传感器漂移：误判接触力，导致抓握物体时施力过大或意外松脱。
- 摄像头遮挡/失效：无法感知障碍物，引发碰撞或跌落楼梯。
- IMU数据异常：错误估计姿态，造成行走失衡或摔倒。

3. 控制器与软件故障
- 控制逻辑错误：步态规划失效导致步态混乱，撞击周围物体。
- 通信延迟/中断：指令不同步引发肢体动作冲突，导致结构损伤。
- AI决策失误：例如强化学习策略在未知环境下生成危险动作（如高速接近人类）。

4. 电源与能源系统故障
- 电池电压骤降：动力不足导致姿态失控，突然倒地。

5. 机械结构故障（关节、连杆）
- 连杆断裂：突然失去支撑力，导致机器人解体或部件飞溅。
- 关节磨损/松动：运动精度下降，引发不可预测的抖动或失衡。

通过上面的分析，我们可以看到从机器人系统的各种安全风险是无处不在的，从而需要从系统设计，管理和安全防范方面去妥善应对这些风险，能够及时监测甚至预测风险，并有效的防范风险。我们认为一个安全子系统的引入是势在必行的，而且是非常有效的一种方案，下面我们将对该安全子系统进行详细介绍。

3. 具身智能机器人安全子系统设计目标

3.1 聚焦物理接触风险

鉴于具身智能机器人与人类共处场景的复杂性和潜在危险，我们的安全子系统将主要精力集中于处理直接的物理接触风险，而将间接风险交予主系统中的应用和安全相关模块去处理。通过这样的设计考虑，我们认为可以实现更聚焦、更高效的安全防护，降低系统复杂度并提高系统可靠性，从而将最紧迫的物理安全风险降至最低。

3.2 参考国际安全标准

为了方便安全系统的设计和评估，目前国际上自动驾驶领域聚焦于不同安全风险的来源产生了多个安全标准，包括功能安全（FuSa）[4]、预期功能安全(SOTIF)[5]以及信息安全。我们的具身智能机器人安全子系统也希望涵盖这些方面，其中主要针对物理安全保障来设计。具体而言，该安全子系统通过对主系统的实时监控和及时干预，作为保障物理安全的最后一道坚固防线，而其中的功能安全和信息安全，均是为了达成物理安全目标，共同构建起一个精简而高效的物理接触安全防护体系。

3.3 基于分级的安全保障

依据 2024 年发布的《人形机器人分类分级应用指南》《具身智能智能化发展阶段分级指南》[6]，人形机器人被划分为 4 个技术等级 L1 - L4，同时具身智能智能化程度被定义为 G1 - G5 五个阶段。而相应的从安全性角度对机器人进行分级，针对不同安全级别的机器人建立相对应的安全监控和保障系统，具有诸多优势。

我们可以将机器人操作任务分为三类：S1（和人不接触）、S2（和人少量接触）、S3（和人紧密接触 / 互动）。对于 S1 类任务，如清理餐桌、整理房间等操作，由于不需要与人近距离接触，安全保障可以采用简单可靠的方法，例如，对任何机器人过于接近人的运动（包括身体和手臂等）都通过安全保障系统进行阻止，让机器人暂时停止，等到距离分开达到安全距离后再恢复运动，以此保证高安全性。而对于 S2、S3 类任务，安全风险检测相对复杂，可能需要引入更为复杂的风险检测算法。当然用于安全风险检测的模型和算法相对主系统中操控任务的模型 / 算法而言尽量要更为简单，有较高的可靠性，尽量避免使用非确定性的算法，这样安全性可以达到较高级别。这种分级方式的好处在于，机器人进入家庭等地方落地服务可能是一个逐步增加功能的过程，刚开始落地应用的人形机器人通常只需执行 S1 级别的操作任务，随着技术和经验的积累，再逐步扩展到更高级别的机器人任务，从而通过大量实际部署来积累宝贵经验，不断完善安全保障系统。这也能尽最大努力防止类似自动驾驶系统安全方面一下子提高到L3以上造成的安全风险。

4. 具身智能机器人安全子系统架构设计与操作模型

4.1 基于“动作单元”的任务执行

为了有效应对日益复杂的任务需求和多变的环境条件，我们提出了一种基于“动作单元”[7] 的操作模型，这是机器人执行任务的基本构建块，代表了一个相对独立、具有特定目标或功能的行为片段。

- **预定义的函数或程序库 ：**例如，一个能够精确控制机械臂移动到特定坐标位置的函数（moveTo (x，y，z)），或者是一个执行特定抓取序列的子程序。这种形式最为传统，易于理解和验证，就像搭积木一样，通过组合这些基本的函数或程序库，可以构建出复杂的机器人任务。

- **特定技能的神经网络模型 ：**比如，一个经过专门训练的神经网络模型，它能够识别并抓取特定类型的物体，或者用于帮助机器人在复杂环境中平稳地进行导航避障。这些模型通过学习大量的数据和模式，使机器人具备了一定的智能和适应性，能够更好地应对多样化任务需求。

- **端到端（E2E）系统中的接口或阶段 ：**在一些先进的 E2E 架构中，如结合快速反应的 System 1 和深思熟虑的 System 2 的系统，“动作单元” 可以代表从一个系统到另一个系统的切换点，或者是 System 1 输出的一个基本行为意图，或者是 System 2 规划出的一个执行阶段。这种设计使得机器人能够在不同层级的智能决策之间灵活转换，提高任务执行的效率和准确性。

无论“动作单元”的内部实现形式如何，每个“动作单元”都应具备可被监控、可被管理，并且定义了其适用操作条件（Operational Design Domain，ODD）的特性。这一特性确保了机器人在执行任务过程中，能够始终处于可控和安全的状态。

基于“动作单元”的操作模型具有诸多优势：

- **模块化与组合 ：**复杂任务可以通过灵活组合不同的动作单元来完成，极大提高了开发过程中的灵活性和效率。例如，一个工业生产中的复杂装配任务，可以分解为多个动作单单元，如pickObject（object_type）、placeAt（location）、

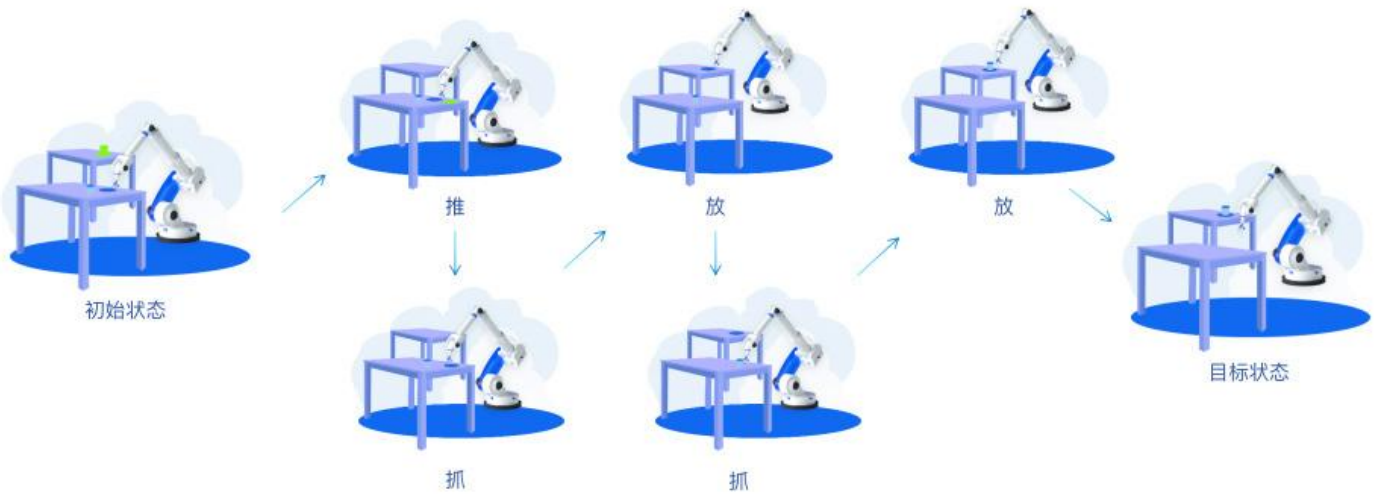


图1 - 动作单元示意图

weldSeam（start_point，end_point）等，通过合理组合这些动作单元，能够快速构建出满足生产需求的机器人工作流程。

- **可测试性与验证 ：** 针对每个动作单元，可以进行独立的测试和验证，从而确保其在特定条件下的功能和安全性。相比于对庞大、单一的系统进行测试，这种方式更加高效、精准，能够及时发现并解决潜在问题，降低开发成本和风险。

- **可重用性 ：** 一旦定义良好的动作单元经过验证，便可以在不同的任务和应用场景中被反复使用。这不仅可以减少重复开发工作，提高资源利用率，还能够保证任务执行的一致性和可靠性。

- **任务规划简化 ：** 任务规划器只需专注于选择和排序合适的动作单元序列，而无需深入到底层控制细节。这使得任务规划过程更加高效、直观，能够更好地适应任务需求的变化和环境的不确定性。

4.2 PMDF 安全架构

受汽车安全理念的启发[8]，我们引入了类似的 PMDF（Primary，Monitor，Decision，Fail Recovery）安全架构，旨在系统性地管理风险，确保机器人系统在各种复杂情况下的安全运行。系统架构图见图2，该架构将带有安全功能的机器人系统按功能划分为四个逻辑上相互独立的模块，分别是具身智能主控系统(P)， 监控系统(M)，安全决策(D)还有故障处理和恢复(F)。而后面这三个模块就是一个安全子系统，也是本白皮书的重点，这个安全子系统在实现上可以是一个单独的硬件如RISC-V，这样可以保障高安全性，而且可以直接放在机器人本体上方便部署。各模块之间协同合作，共同构建起坚实的安全防线。除此之外，如果要增加安全性，可以引入可选的外部安全系统，这个系统可以部署在机器人外部(环境中)，这样就形成了一个多级机器人安全系统。外部安全系统这部分我们在这里不做具体讨论。下面我们对这个架构的主要部分也就是PDMF各个模块和它们之间的关系进行介绍。

- **P（主控模块） ：** 主控模块承担着机器人主要任务规划和执行逻辑的重任。它依据高级目标，例如 “将 A 物体搬运到 B 点”，生成一系列 “动作单元” 的序列或指令。P 模块专注于 “做什么” 和 “如何做”，即负责正常流程下的任务执行。虽然 P 模块本身可能具备一定的安全意识，例如进行基本的路径规划避障，但其核心目标始终是完成既定任务，因此主要的安全保障依赖于安全子系统各个模块的协作。

- **M（监控模块）：** 监控模块是系统的 “守护者”，其核心职责在于持续、独立地监控系统状态和外部环境，是实现功能安全（FuSa）和预期功能安全（SOTIF）的关键模块。它具备以下功能：
  - 行为监控 ： 实时监测当前执行的动作单元是否符合预期。例如，检查机械臂的实际速度、轨迹、末端受力等参数是否在安全和预期的范围内；对于导航机器人，监控其实际位置是否偏离了计划路径。
  - 环境与 ODD 监控 ： 利用各类传感器，如摄像头、激光雷达、力传感器、麦克风等，感知周围环境，并评估当前环境是否满足待执行或正在执行的 “动作单元” 的操作设计域（ODD）。例如，在执行一个快速移动的动作

单元之前，M 模块需要确认机器人附近没有障碍物或人；在执行一个需要精确视觉定位的动作单元前，M 模块需要确保光照条件符合要求，以保障视觉系统的准确性。

- 系统状态监控： 监测 P、D、F 以及 M 自身模块的健康状态和功能是否正常。这包括心跳检测、关键进程响应、传感器数据有效性检查等，是功能安全（FuSa）的核心内容。通过这些监控功能，M 模块能够及时发现系统中的异常情况，并迅速向 D 模块发送警告或错误信号。

- **D（决策模块）：**决策模块作为安全决策的核心枢纽，其核心职责是根据来自 P 和 M 的信息做出最终的执行决策。在正常模式下，当 M 模块没有报告异常且系统状态正常时，D 模块会透明地将 P 模块发来的动作单元指令转发给底层的执行控制器，确保任务的顺利执行。然而，一旦接收到来自 M 模块的警告或错误信号，D 模块将立即切换到异常模式：
  - 阻止 / 否决： 阻止 P 模块当前或后续可能导致危险的指令被执行，从而避免潜在的安全事故。
  - 激活 F： 根据 M 模块报告的异常类型和严重程度，激活 F（Fail Recovery）模块执行相应的安全策略，例如使机器人进入安全停止状态或启动其他故障恢复措施。
  - 仲裁： 在某些特殊情况下，P 和 M 模块可能提供冲突的信息，例如 P 请求机器人移动，但 M 检测到前方有障碍物。此时，D 模块将依据预设的安全规则进行仲裁，始终将安全放在首位，做出合理的决策。

  为了确保决策的公正性和可靠性，D 模块应设计为尽可能独立，其决策逻辑不应被 P 模块轻易覆盖。

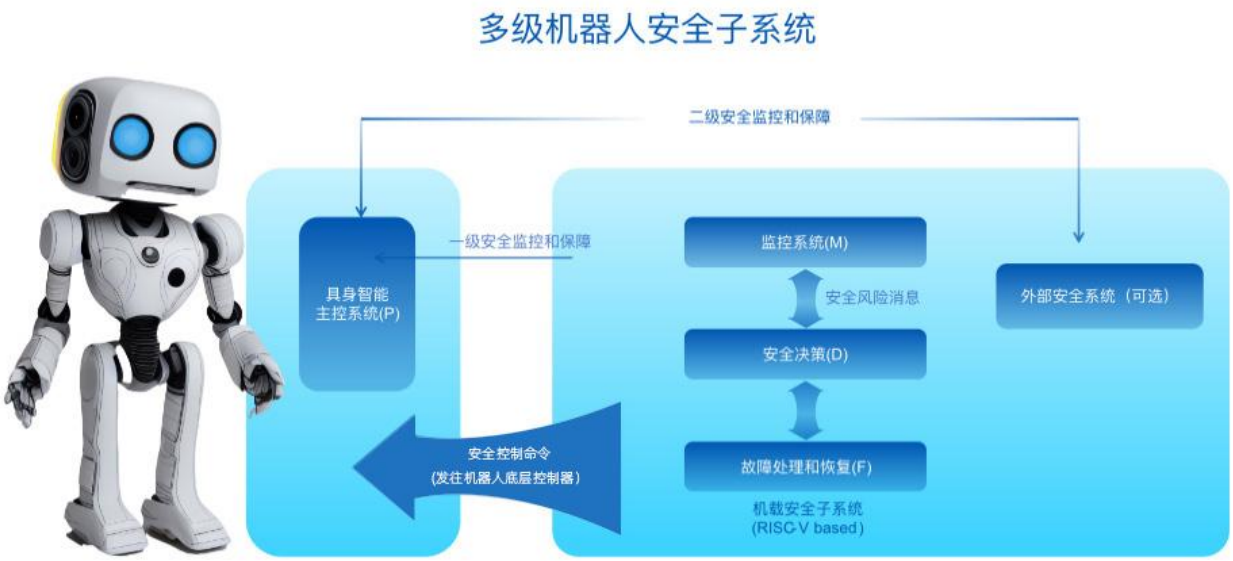


图2 - 人形机器人PMDF安全系统架构图

- **F（故障恢复模块）** ：故障恢复模块的核心职责是执行预定义的、使系统进入最小风险状态（Minimal Risk Condition，MRC）的策略。这些策略通常具有简单、可靠且经过充分验证的特点，具体策略会根据机器人类型和应用场景有所不同。例如，故障恢复策略可能包括：
  - 安全停止： 立即停止机器人的所有运动，这是最常见的故障恢复措施之一，能够在紧急情况下迅速消除潜在危险。
  - 进入怠速 / 安全模式： 保持机器人通电但停止任务执行，等待进一步的指令或人工干预。这为操作人员提供了足够的时间来评估情况并采取相应措施。
  - 执行特定恢复动作： 根据故障类型，执行一些特定的恢复动作，如松开夹爪以释放物体、将机械臂移动到安全的停靠位置等，以减少故障对机器人和周围环境的影响。
  - 发出警报： 通过声音、灯光或网络通知等方式提醒操作员或用户，告知他们机器人系统出现了故障，需要及时处理。

  F 模块仅在接收到 D 模块明确的激活指令时才执行操作，确保故障恢复过程的有序性和可控性。

4.3 模块交互与信息流

PMDF 架构的有效性依赖于各模块之间清晰、可靠的信息流。图3是PMDF架构信息流示意图。

（P 将指令发送给 D，M 将监控信息（状态/环境/ODD）发送给 D，D 根据 M 的信息决定是转发 P 的指令给执行器，还是激活 F 执行恢复策略。）

以下是模块交互与信息流的具体过程：

- **任务启动：** P模块根据任务需求，规划出第一个或下一系列动作单元指令，为任务的执行做好准备。

- **指令传递：** P模块将规划好的动作单元指令发送给 D 模块，并等待 D 模块的审批和转发。

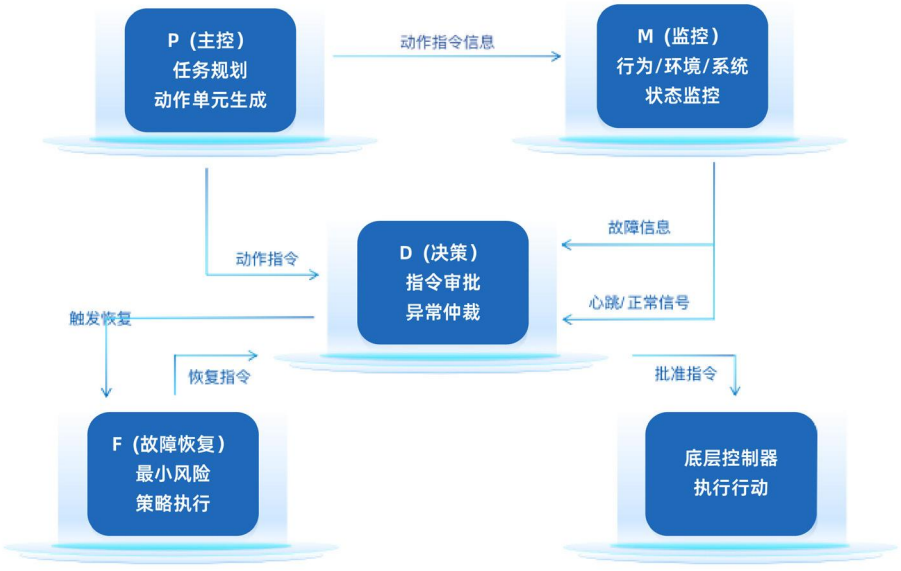


图3 - PMDF 架构信息流示意图

- **状态 / 环境监控：** 与此同时，M 模块持续不断地监控系统的实时行为、外部环境以及操作设计域（ODD）等信息，并对系统各模块的健康状态进行监测。

- **监控报告：**
  - 正常情况：如果一切正常，M 模块会周期性地向 D 模块发送 “一切正常” 的心跳信号，告知 D 模块系统处于稳定状态。
  - 异常情况：一旦 M 模块检测到任何不安全或异常情况，如行为偏差、环境不满足 ODD、系统模块故障等，它会立即向 D 模块发送具体的警告或错误信息，详细描述异常的类型、位置和严重程度等关键信息，例如 “ODD 不满足：检测到近距离障碍物” “P 模块无响应” “关节 3 电流过载” 等。

- **决策与执行：**
  - 正常模式：当 D 模块接收到 P 模块的指令以及 M 模块的正常信号后，会立即批准该指令，并将其转发给底层的执行控制器，确保任务顺利执行。
  - 异常模式：如果 D 模块接收到 P 模块的指令，但同时收到 M 模块的异常信号，它会果断阻止 P 模块的指令，并根据 M 模块提供的异常信息激活 F 模块，执行相应的恢复策略，以保障系统安全。
  - 无指令异常：在 D 模块未收到 P 模块指令的情况下，若收到 M 模块的异常信号（例如系统出现严重故障），D 模块会直接激活 F 模块，执行故障恢复策略，迅速将系统带入安全状态。

- **故障恢复：** F 模块在接收到 D 模块的激活指令后，立即执行预设的最小风险策略，如安全停止、进入怠速模式等，以最大程度减少故障对系统和周围环境的影响。

通过这种架构，我们将任务执行（P）、安全监控（M）、安全决策（D）以及故障响应（F）分离开来，实现了各模块职责的清晰划分。这种设计不仅有助于构建更加健壮、可靠的机器人系统，还能够提高系统的可维护性和可扩展性。同时，对 “动作单元” 的灵活定义使该架构能够适应不同的机器人技术和实现方式，具有广泛的适用性和良好的适应性。

4.4 安全理念的映射与整合

我们提出的PMDF架构能够与功能安全（FuSa）、预期功能安全（SOTIF）以及信息安全等关键安全理念深度融合，共同构筑起机器人系统的坚固安全防线。

在功能安全方面，PMDF架构通过其精巧设计保障系统基础可靠性。关键模块采用冗余设计与定期自检机制，如同为系统配备多重 “保险丝”，即使个别部件出现故障，也能确保整体安全功能不受影响。系统启动时进行严格自检，如同对机器人进行一次全面 “体检”，只有所有关键安全组件 “健康达标”，系统才能投入运行，从源头避免带病作业，确保机器人在执行任务过程中可靠运行，是功能安全落地的有力载体。

对于预期功能安全（SOTIF），PMDF架构更是发挥重要作用。其能实时评估机器人执行任务时的环境条件，精准判断是否符合动作单元预设的操作设计域（ODD），当环境变化超出机器人应对能力时，迅速采取措施保障安全。例如在能见度降低时，及时调整行动策

略或触发安全机制，有效应对功能局限与环境挑战，是实现SOTIF的关键支撑。

在信息安全领域，PMDF架构全方位守护机器人系统免受外部威胁。通过加密通信链路、强化身份认证以及实时入侵检测等手段，为机器人打造坚不可摧的网络安全屏障。既能防止恶意攻击者窃取敏感数据，又能抵御对机器人控制权的恶意篡夺，从源头肃清安全隐患，为机器人在复杂网络环境中的安全运行保驾护航。通过这种紧密映射与深度融合，PMDF架构将多种安全理念贯穿于机器人系统的设计、运行与维护全过程，实现安全性能的全面提升，为机器人在复杂多变环境中稳定、可靠运行奠定坚实基础。

4.5一个硬件实现实例

图4是 一个硬件实现的实际例子，这里我们采用了RISC-V安全芯片作为MDF系统的实际硬件平台，从而可以在安全性和成本控制方面达到一个较好的平衡。该安全芯片在设计上支持了功能安全，并且有较好的信息安全功能，这样为整个的安全子系统提供了坚实的基础。前面提到的MDF系统都可以在该芯片上通过虚拟机和软件模块结合的方式来实现很高的安全性，而由于传感器和相应健康和决策算法都采用比较可靠且计算量可控的算法，也能保障效率还有实时性。这里的主计算单元可以有多种选择，比如目前比较采用较多的Intel CPU加Nvidia Orin的组合，也有新出现的Intel CPU加英特尔独立显卡B580的组合，各个厂家可以根据自己的情况进行选择，只要其系统实现中和安全子系统进行相应的接口就可以通过或安全子系统来保障机器人整体的物理安全。

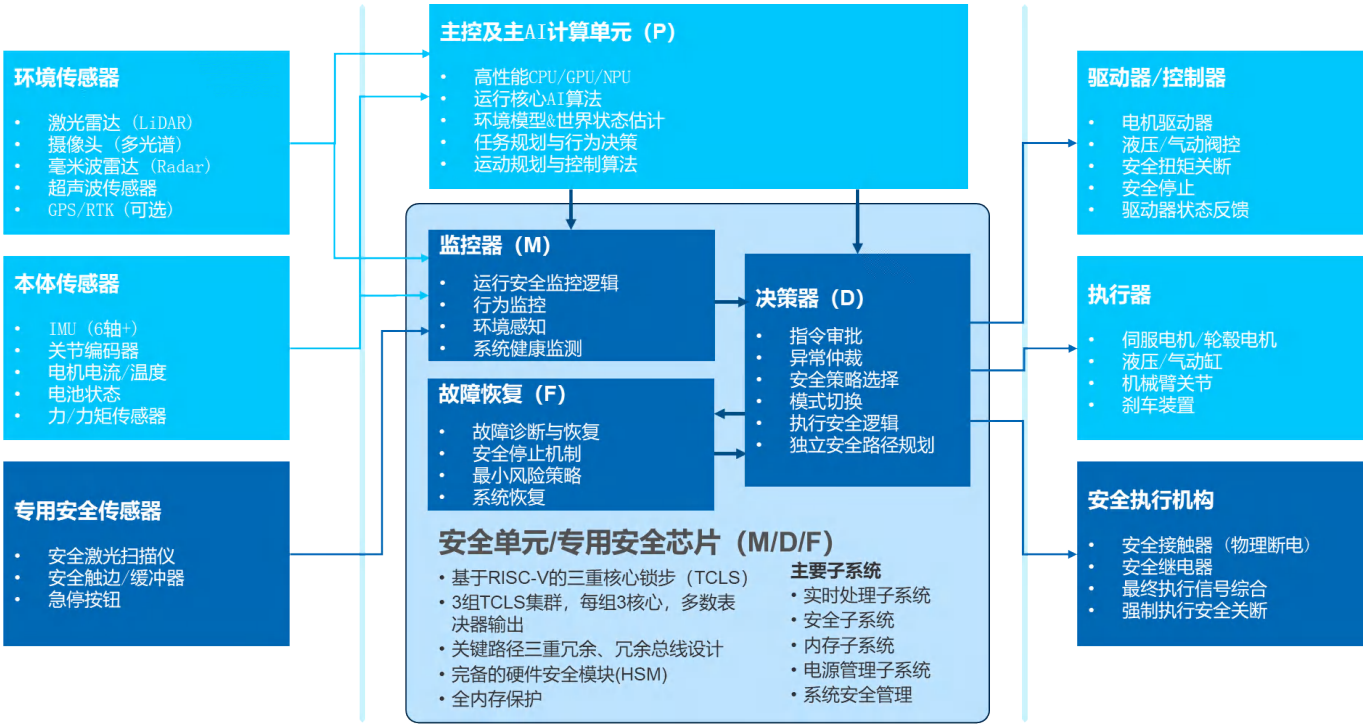


图4 - 安全子系统的一个硬件实现实例

5. 总结

在人工智能与机器人技术飞速发展的浪潮中，具身智能机器人正逐步走进我们的生活和工作场景，展现出巨大的应用潜力。然而，安全问题始终是人形机器人发展道路上不可忽视的重要挑战。本文深入探讨了具身智能机器人的物理安全问题，分析了其面临的各种风险以及风险产生的原因，并提出了针对性的安全子系统设计目标。我们引入了基于“动作单元”的操作模型和 PMDF 安全架构，旨在为机器人系统提供全方位、多层次的安全保障。

尽管目前具身智能机器人的安全技术已经取得了一定的进展，但仍然有许多工作需要进一步开展。在未来几年中，我们将继续深入研究，不断完善安全子系统的功能和性能。我们相信，通过与广大生态系统伙伴们的共同努力与协作，我们能够携手克服技术难题，推动个人服务机器人更快地进入家庭，为人们提供更加智能、便捷、安全的服务，实现具身智能机器人与人类和谐共处的美好愿景。

参考文献

[1] 中国国家标准（2019）GB/T 38244-2019，机器人安全总则。北京: 中国标准出版社, 2019

[2] Vasic, Milos, and Aude Billard. "Safety issues in human-robot interactions." 2013 ieee international conference on robotics and automation. IEEE, 2013.

[3] The Humanoid 100: Mapping the Humanoid Robot Value Chain, Morgan Stanley Research Report, 2025.
https://advisor.morganstanley.com/john.howard/documents/field/j/jo/john-howard/The_Humanoid_100_-_Mapping_the_Humanoid_Robot_Value_Chain.pdf

[4] International Organization for Standardization. (2018). ISO 26262:2018 Road vehicles – Functional safety. Geneva, Switzerland: ISO.

[5] International Organization for Standardization. (2022). ISO 21448:2022 Road vehicles – Safety of the intended functionality. Geneva, Switzerland: ISO.

[6] 全国首批人形机器人具身智能标准在浦东发布,
https://www.shanghai.gov.cn/nw15343/20241030/e0b86ec27fb84890b9badfbd7d786507.html

[7] Strudel, Robin, et al. "Learning to combine primitive skills: A step towards versatile robotic manipulation §." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.

[8] An Architecture for Safe Driving Automation，Hermann Kopetz，Principles of Systems Design，2022: 61-84

附录：PMDF 架构技术细节

关键模块的可靠性设计

- **冗余与自检 ：** 对于 PMDF 架构中的核心安全模块，如 M（监控模块）、D（决策模块）、F（故障恢复模块），采用硬件冗余设计是一种常见的提升可靠性方法。例如，采用双通道处理器，当主处理器出现故障时，备用处理器能够迅速接管工作，确保系统的不间断运行；对于关键传感器，如激光雷达、摄像头等，配置冗余传感器，当主传感器数据异常或故障时，切换到备用传感器，从而提高系统的可靠性和容错能力。同时，这些模块还应具备启动自检和周期性在线自检功能。在系统启动时，对模块的硬件和软件进行全面检查，确保其处于良好的初始状态；在系统运行过程中，定期进行自检，及时发现并报告潜在的故障隐患。例如，D 模块可以设计为双核锁步架构，两个处理器同时执行相同的任务，并相互比较结果，一旦发现结果不一致，立即触发故障处理机制，从而有效提高决策模块的可靠性和安全性。
- **M 模块的自监控 ：** M 模块不仅要负责监控其他模块和系统状态，还需要对自身的监控功能进行自我监控。这包括对传感器输入数据的有效性检查、算法运行状态的监测等。为了确保 M 模块自身的可靠性，其用于自监控和核心安全判断的算法逻辑应尽可能采用确定性高、可形式化验证的方法。例如，基于规则的逻辑判断、简单的数学模型等，这些方法具有较高的可验证性和稳定性，能够有效避免因算法本身的不确定性而导致的监控失效。同时，M 模块应具备自我修复能力，当检测到自身出现轻微故障时，能够自动采取措施进行修复，如重新初始化传感器、重启相关进程等，以确保其监控功能的持续有效性。

系统启动安全检查

在机器人系统启动过程中，必须执行严格的自检程序，全面检查 P、M、D、F 各模块及其相关组件的初始化状态和功能完整性。具体检查内容包括但不限于以下几个方面：

- **模块初始化检查 ：** 确认各模块的软件和硬件是否成功完成初始化操作。例如，检查 P 模块的任务规划算法是否加载正确、参数设置是否合理；M 模块的传感器驱动程序是否正常启动、数据采集功能是否可用；D 模块的决策逻辑是否初始化完成、通信接口是否正常工作；F 模块的故障恢复策略是否就绪等。
- **传感器和执行器状态检测 ：** 对系统中的各类传感器和执行器进行全面检查。对于传感器，检测其供电是否正常、通信链路是否通畅、输出数据是否符合预期格式和范围等。例如，检查摄像头是否能够正常采集图像、图像数据是否清晰完整；激光雷达是否能够正常旋转并输出有效的点云数据；力传感器是否能够准确测量力的大小和方向等。对于执行器，如电机驱动器、机械臂关节等，检测其是否能够正常响应控制信号、运行是否平稳、是否存在过载或故障等情况。
- **通信链路检查 ：** 验证系统内部各模块之间的通信链路以及机器人与外部设备（如控制中心、云平台等）的通信接口是否正常工作。检查数据传输的准确性、完整性和实时性，确保各模块之间能够及时、准确地交换信息。

只有当所有关键安全组件均通过上述严格的自检程序，确认处于正常工作状态时，机器人系统才能顺利进入操作模式，开始执行预定任务。如果在启动自检过程中发现任何问题，系统将禁止进入操作模式，并立即触发警报机制，提示维护人员进行检查和修复，从而有效避免因系统故障而导致的安全事故。

故障检测与报告机制

- **M 模块的核心职责：** M 模块作为系统监控的核心，承担着持续监测 P、D、F 模块以及整个系统状态的重要职责。它通过实时收集和分析各模块的心跳信号、响应时间、错误码等关键信息，及时发现模块是否存在故障。例如，M 模块可以通过设置定时器来监测各模块的响应时间，若某一模块的响应时间超出设定的合理阈值，则判定该模块可能出现响应超时故障；同时，通过解析模块发送的错误码，能够快速识别出硬件故障、软件异常、通信错误等各种问题，从而为系统的安全运行提供及时的预警和决策支持。
- **传感器 / 执行器诊断：** M 模块还需对关键传感器和执行器进行全面的诊断。对于传感器，不仅要检查其输出数据的有效性，如是否存在信号丢失、数据异常波动、测量值超出合理范围等情况，还要对其性能进行评估，如传感器的精度、分辨率、重复性等是否符合要求。对于执行器，监测其运行状态，如电机的转速、扭矩、温度等参数是否正常，驱动器是否存在故障报警信号等。通过这些细致的诊断措施，M 模块能够及时发现传感器和执行器的潜在故障，确保系统的感知和执行能力不受影响。
- **故障信号传递：** 一旦 M 模块检测到任何故障信息，必须迅速、准确地将故障类型、位置、严重程度等详细信息传递给 D 模块。为了保证故障信号传递的可靠性，系统应采用冗余的通信链路和容错机制。例如，采用双通道通信机制，当主通道出现故障时，自动切换到备用通道，确保故障信息能够及时传达给 D 模块，从而为 D 模块做出正确的决策提供充足的时间和准确的信息依据。

定义安全状态 (Safe State) 与故障响应

- **最小风险条件 (MRC)：** 针对机器人系统可能遇到的各种关键故障场景，应明确且详细地定义系统应进入的最小风险状态。例如，在机器人机械臂关节发生故障时，MRC 可能是将机械臂以最缓慢的速度移动到预先设定的安全停靠位置，并锁定关节，防止其因失控而产生意外动作，同时切断动力源，确保机械臂完全停止运动；在导航系统出现故障时，MRC 是使机器人立即停止所有运动，并通过声音、灯光等多种方式发出警报，提醒周围人员注意安全，同时记录故障信息，等待维护人员的检查和修复。
- **D 模块的故障处理：** 当 D 模块接收到 M 模块报告的严重故障信息后，必须按照预设的优先级和流程，迅速触发 F 模块执行相应的 MRC。在触发过程中，D 模块需要对故障信息进行快速分析和处理，确保故障处理的及时性和有效性。同时，D 模块应记录故障处理过程中的关键信息，如故障发生的时间、类型、处理措施、结果等，以便后续对故障进行深入分析和改进系统设计，提高系统的可靠性和安全性。此外，D 模块还应具备一定的自诊断能力，当自身出现故障时，能够及时发现并采取措施，如切换到备用决策逻辑或触发更高层级的故障处理机制，确保系统的安全决策功能不受影响。

动作单元的操作设计域 (ODD) 定义

- **明确边界：** 为每个 “动作单元” 严格定义其安全可靠运行的条件，是确保机器人系统安全性的关键环节。这些条件涵盖了环境因素（如光照范围、天气条件、地面类型、允许的动态障碍物密度/速度）、系统状态（如自身速度、负载）以及与其他实体的交互规则（如与人的安全距离）。例如，对于一个 “切割物体” 的动作单元，其 ODD 可以详细定义为：环境光照强度必须在 300 - 1000lux 之间，以保证视觉系统能够准确识别切割对象和周围环境；周围 1 米范围内不得有人或其他动物，避免切割过程中产生的碎片或意外动作对其造成伤害；机器人自身定位精度需高于 ±5mm，以确保切割工具能够精确到达目标位置；切割工具必须处于良好的工作状态，

如刀片无损坏、磨损等。通过这些明确的边界条件，为动作单元的执行提供了清晰的安全约束，确保其在设计范围内的可靠运行。

- **量化与可测性 ：** 在定义 ODD 时，应尽可能采用量化的方式，并确保这些条件能够通过传感器进行实时测量和验证。例如，对于环境光照强度，可使用专业的光照传感器进行实时监测，并将其测量值与 ODD 中设定的范围进行比较；对于动态障碍物密度和速度，可结合激光雷达和视觉传感器的数据进行计算和评估，通过特定的算法模型将传感器数据转换为可量化的指标，从而实现对 ODD 条件的准确判断。这种量化与可测性的设计不仅提高了 ODD 的可操作性，还为系统的自动化监控和决策提供了可靠的数据支持。

对监控算法的要求与考量

- **优先采用传统算法 ：** 在 M 模块的 ODD 验证过程中，优先选择那些具有较高确定性和可解释性的传统算法，如基于几何规则的障碍物检测算法、基于明确阈值的传感器数据检查算法、逻辑规则引擎等。这些算法通常具有较为成熟的理论基础和广泛的应用经验，便于进行形式化验证和 exhaustive testing，从而确保其在安全关键应用中的可靠性和稳定性。例如，利用几何规则检测障碍物的形状、大小和位置，通过设定阈值判断传感器数据是否异常，以及运用逻辑规则引擎对多个传感器数据进行融合和推理，能够准确地确定环境是否满足 ODD 要求，为机器人的安全运行提供有力保障。
- **谨慎使用神经网络 ：** 在某些复杂场景下，若必须使用神经网络（如对复杂物体进行识别），则需要采取一系列额外的措施来约束其行为并提高其可靠性。例如，通过设置保守的安全边界，对神经网络的输出结果进行严格限制，当输出结果超出安全边界时，触发相应的安全机制；同时，对神经网络的输出进行独立的安全校验，可采用一个简单的、可验证的算法对网络输出进行再次验证，确保输出结果符合安全要求。此外，还可以通过增加神经网络的训练数据量、优化网络结构、采用模型融合等方法，提高其在复杂环境下的性能和稳定性，降低其不确定性对系统安全的影响。
- **强调可测试性和可复现性 ：** 无论采用何种算法，都必须确保其行为具有可测试性和可复现性，并且明确其性能指标和失效模式。在算法设计阶段，应充分考虑如何对其进行有效的测试和验证，设计各种测试用例，包括正常场景、边界条件和异常情况，以全面评估算法的性能和可靠性。同时，记录算法在不同场景下的表现和输出结果，分析其性能指标（如检测准确率、误报率、漏报率等）以及失效模式（如在特定光照条件下识别错误、对特定形状障碍物检测失败等），并根据分析结果对算法进行优化和改进。在实际应用中，通过对算法输入相同的测试数据，能够获得相同或相似的输出结果，从而保证算法的可复现性，为系统的安全运行提供可靠的算法支持。

触发条件 (Triggering Conditions) 分析与管理

- **识别风险场景 ：** 系统性地分析和识别可能导致机器人偏离安全 ODD 或功能表现下降的各种环境因素、系统行为或交互情况，是预防安全事故的关键。例如，当传感器被遮挡时，机器人可能无法准确感知周围环境，从而导致决策失误；强光干扰视觉系统，会使物体识别精度下降；遇到未训练过的物体，机器人的识别算法可能无法正确分类，进而采取错误的行动；人突然闯入机器人的工作区域，可能与机器人发生碰撞或干扰其正常运行等。通过对这些风险场景的深入分析，可以提前制定相应的应对策略，提高系统的安全性和鲁棒性。
- **M 模块的检测 ：** M 模块需要具备强大的检测能力，能够实时监测和识别上述潜在的触发条件。这要求 M 模块充分利用各种传感器资源，结合先进的信号处理和模式识别算法，对环境和系统

状态进行全方位的感知和分析。例如，利用摄像头的图像处理算法检测强光干扰，通过激光雷达的点云数据判断是否有物体或人突然闯入机器人的安全距离范围内；对传感器数据进行实时分析，判断是否存在遮挡、损坏或其他异常情况，从而及时发现风险场景的出现，并向 D 模块发出警告信号。

功能局限性认知与应对

- **明确能力边界 ：** 如实记录传感器、感知算法和决策模型的局限性，是确保机器人系统安全运行的重要前提。例如，基于视觉的障碍物检测算法在浓雾天气下其检测距离可能大幅缩短，漏检率显著增加；对于某些小型或形状奇特的物体，感知算法可能无法准确识别其类别和位置；决策模型在处理高度复杂的动态环境时，可能存在反应迟缓或决策失误的情况等。通过对这些局限性的深入分析和明确记录，能够帮助系统更好地了解自身的能力范围，避免在不适宜的条件下执行任务，从而降低安全风险。
- **安全裕度 ：** 在定义 ODD 和设计监控策略时，应当充分考虑上述局限性，并预留足够的安全裕度。例如，在定义机器人与人的安全距离时，不仅要考虑正常情况下的运动学模型和反应时间，还需预留一定的裕度，以应对传感器误差、系统延迟或人的突然动作等情况。通过增加安全裕度，能够为系统提供额外的缓冲空间，提高系统在面对不确定性和异常情况时的安全性和可靠性。

测试与验证 (V&V)

- **场景库 ：** 建立一个全面且丰富的测试用例库，涵盖各种安全关键场景、边界条件和潜在触发条件，是评估和验证机器人系统安全性的基础。该场景库应包括仿真场景和实测场景，以充分模拟机器人在不同环境和任务下的运行情况。例如，针对户外机器人，可设计强光、弱光、雨天、雪天等不同光照和天气条件下的仿真场景，以及在不同地形、障碍物密度和人流量下的实测场景；针对服务机器人，可构建家庭、医院、商场等不同环境下的测试场景，模拟各种可能遇到的情况，如人突然倒地、物体掉落、地面湿滑等。这些测试场景为系统的测试与验证提供了丰富的测试环境，有助于全面评估系统的安全性能。
- **验证 M 的 ODD 判断 ：** 重点测试 M 模块在各种条件下判断 ODD 是否满足的准确性和及时性，尤其是算法在边界条件下的表现和鲁棒性。通过大量的测试用例，评估 M 模块在不同环境下的 ODD 判断能力，记录其判断结果与实际情况的偏差，分析偏差产生的原因，如传感器精度不足、算法模型局限等，并据此对 M 模块进行优化和改进，提高其在边界条件下的判断准确性和可靠性。例如，在模拟强光干扰的场景下，测试 M 模块是否能够准确判断视觉系统的 ODD 是否满足，从而为 D 模块做出正确的决策提供依据。
- **验证整体响应 ：** 测试从 M 模块检测到 ODD 不满足到 D 模块做出决策，再到 F 模块执行安全策略的整个响应链条的完整性和有效性。通过模拟各种故障场景和风险情况，检查各模块之间的通信是否正常、决策是否正确、安全策略是否执行到位，确保整个响应过程能够快速、准确地保障机器人和人的安全。例如，在测试中故意遮挡传感器，观察 M 模块是否及时检测到 ODD 不满足，D 模块是否正确阻止 P 模块的危险动作，并激活 F 模块执行相应的安全策略，如减速、停车或发出警报等，从而验证整个系统的安全响应机制的有效性。

不满足 ODD 时的响应机制

- **M -> D -> F/P ：** 当 M 模块检测到当前或即将进入的环境不满足动作单元的 ODD 时，立即向 D 模块发送包含详细信息的通知，包括不满足的条件、可能的风险以及建议的应对措施等，以便D

模块能够快速做出决策。

- D 的决策：D 模块依据 M 模块提供的信息，综合考虑系统的整体状态和安全需求，可采取多种决策措施。例如，若当前环境光照不足，无法满足视觉识别动作单元的 ODD，D 模块可阻止 P 模块执行该动作单元，并指示 P 模块切换到使用红外或超声波传感器的备用导航动作单元，以适应当前环境条件；若环境存在高风险因素（如附近有火源），D 模块可直接触发 F 模块执行最小风险策略，如停止机器人运动、关闭相关功能模块等，迅速将系统带入安全状态，避免潜在的危险发生。

致谢：

在白皮书的撰写过程中，我们得到了众多同事和朋友的无私帮助与支持。特别感谢英特尔中国研究院参与具身智能相关项目的各位同事，他们在机器人项目中的丰富经验和专业知识为我们提供了宝贵的指导。同时，也要感谢那些在应用和技术需求调研过程中积极参与讨论并提出建设性意见的同仁和朋友们，正是他们的热情参与和智慧贡献，使得这份白皮书得以不断完善。