中国人工智能系列白皮书

—-具身智能（2026 版）

中国人工智能学会

二○二六年四月

编委会

目　　录

第一章　具身智能的概念与内涵

具身智能作为人工智能领域的一个重要研究方向，专注于智能体通过物理本体与外界环境的互动来实现智能的理论与技术研究，涵盖环境感知、记忆推理、对话交互、自主学习、决策规划、动作执行等综合性技术，从而在真实物理世界中展示出类人的智能行为。相比于静态、离身的人工智能，具身智能具有涉身性、情境性、主动性和交互性等特点。具身智能兼具多技术融合与多学科交叉特性，与计算机科学、机器人学、神经科学、认知科学等不同领域都紧密相关，其研究范畴、研究范式，内涵外延也在不断发展中。具身智能近年来得到了学术界、产业界的大量关注，被认为是人工智能的下一个爆发点，是人工智能走向物理世界的核心关键，在国计民生等各方面都有重大潜在应用价值。

1.1　具身智能发展历史

具身智能的演进历程可追溯至 20 世纪 50 年代，其理论源自英国杰出的计算机科学家阿兰 · 图灵（Alan Turing）的深刻洞见。1950 年，图灵在其具有划时代意义的论文《计算机器与智能》（Computing Machinery and Intelligence）中，首次构想了一种能够与环境进行动态交互、具备自我学习能力的智能实体。该智能体被设想为能够像人类一样感知外界环境、自主规划行动路径、做出决策，并具备高效执行任务的能力，这一构想被视为具身智能（Embodied Intelligence）的初步理论框架。

步入 20 世纪 80 年代，随着人工智能研究的不断深入，行为主义 AI 学派开始崭露头角，其中罗德尼 · 布鲁克斯（Rodney Brooks）等学者的研究尤为突出。他们强调通过感知与动作的紧密协同，设计能够与环境进行有效交互的智能机器。这一时期的“具身”机器人实验主要聚焦于利用逻辑规则算法与机器人硬件的结合，以实现特定的应用功能。尽管这些实验尚处于初步探索阶段，但它们为具身智能的发展奠定了重要基础。

随着技术的不断积累与创新，具身智能迎来了快速发展的黄金时期。深度学习（Deep Learning）、强化学习（Reinforcement Learning）等先进算法模型的涌现，为具身智能提供了强大的技术支持。这些算法模型使机器人能够更好地理解和处理复杂的环境信息，从而实现更加智能和灵活的行为。同时，传感器与执行器等硬件技术的

不断进步，也显著提升了机器人的感知敏锐度和行动精准度。在这一阶段，“具身”机器人技术取得了显著进展，不仅在仿生机器人研发方面取得了重要突破，还在“人工智能 + 机器人”的智能化融合上迈出了决定性步伐。例如，特斯拉的人形机器人 Optimus 通过先进的视觉-语言-动作模型以及精确的电机控制技术，实现了智能、拟人的交互，展示了具身智能在机器人领域的巨大潜力。

近年来，随着大语言模型（Large Language Models，LLMs）的兴起，具身智能的发展迎来了新的高潮。大模型凭借其深厚的通用知识库和智能涌现能力，为机器人提供了更高层次的智能感知、自主决策和拟人化交互能力。谷歌 DeepMind 推出的 RT 系列机器人，尤其是 RT-H 版本，通过创新的任务分解与语言指令转化策略，实现了任务执行的高精度与高效率，进一步推动了具身智能在复杂任务处理方面的能力。

此外，Meta AI 发布的 CortexBench 视觉评估基准以及专为具身智能设计的 VC-1 视觉模型，为具身智能的标准化评估与模型优化提供了重要工具。英伟达（NVIDIA）作为 GPU 和 AI 计算领域的领导者，在具身智能领域同样做出了显著贡献。他们推出了 GR00T 人形机器人基础模型及 Jetson Thor 新款人形机器人计算机，并对 Isaac 机器人开发平台进行了全面升级，为机器人技术的持续创新提供了有力支撑。

1.2 具身智能多学科交叉特性

具身智能的思想和研究跨越了多个学科，包括哲学、认知与神经科学、计算机科学、机器人学等，体现了显著的多学科交叉特性。

古希腊的亚里士多德就开始思考心灵与身体的关系。他在《论灵魂》中认为，心灵是生命体的本质和形式，赋予生物以感知、思考和运动的能力。20 世纪 80 年代，认知科学家发展了具身认知理论，认为认知过程不仅仅是大脑的内部活动，而是身体在与环境交互和耦合中产生的。神经科学对镜像神经元的研究发现大脑可以表征其他动物的行为，进一步强化了具身智能在群体交互中的作用。机器人学家通过构建智能机器人发现智能行为可以通过智能体与环境的直接交互实现，而不需要复杂的内部表征。这些研究推动了计算机科学家从感知行动整合的角度研究具身认知的信息映射过程。复杂系统领域的研究者则从演化和信息论的角度指出，智能体的行为可以看作是一个状态随时间演化的复杂动力系统，在信息最优化的原则下进行自组织学习，产生与环境交互的探索行为。近期的一些研究则从强化学习的角度发现，环境的复杂性促进智能形态的进化和代际传递。

具身智能的多学科交叉特性是其研究和发展的重要驱动力。通过跨学科的交叉融合，具身智能不仅推动了人工智能的理论创新，还为解决现实问题提供了新的技术手段。

1.3 具身虚实结合现状

近年来，具身智能领域出现了几种虚拟与现实结合的范式，如图 1-1所示。由于在真实世界中采集专家示范动作序列的时间成本与技术要求较高，直接将虚拟环境中训练得到的策略迁移到真实世界部署会面临严重的“虚拟-现实鸿沟（Sim-to-Real Gap）”。因此，一系列方法应运而生，旨在将虚拟与现实结合，尽可能弥合这一鸿沟。

真实感强化[1–6] 借助高真实感渲染的 3D Gaussian Splatting 等表示方法构建场景和智能体在虚拟环境中的数字孪生，通过增强模拟环境的真实感，将虚拟环境中的专家示范渲染成真实世界的样子，进而让具身智能进行模仿学习，以解决获取真实数据过程中高时间成本和技术成本的问题，同时实现有效的虚拟-现实策略迁移。

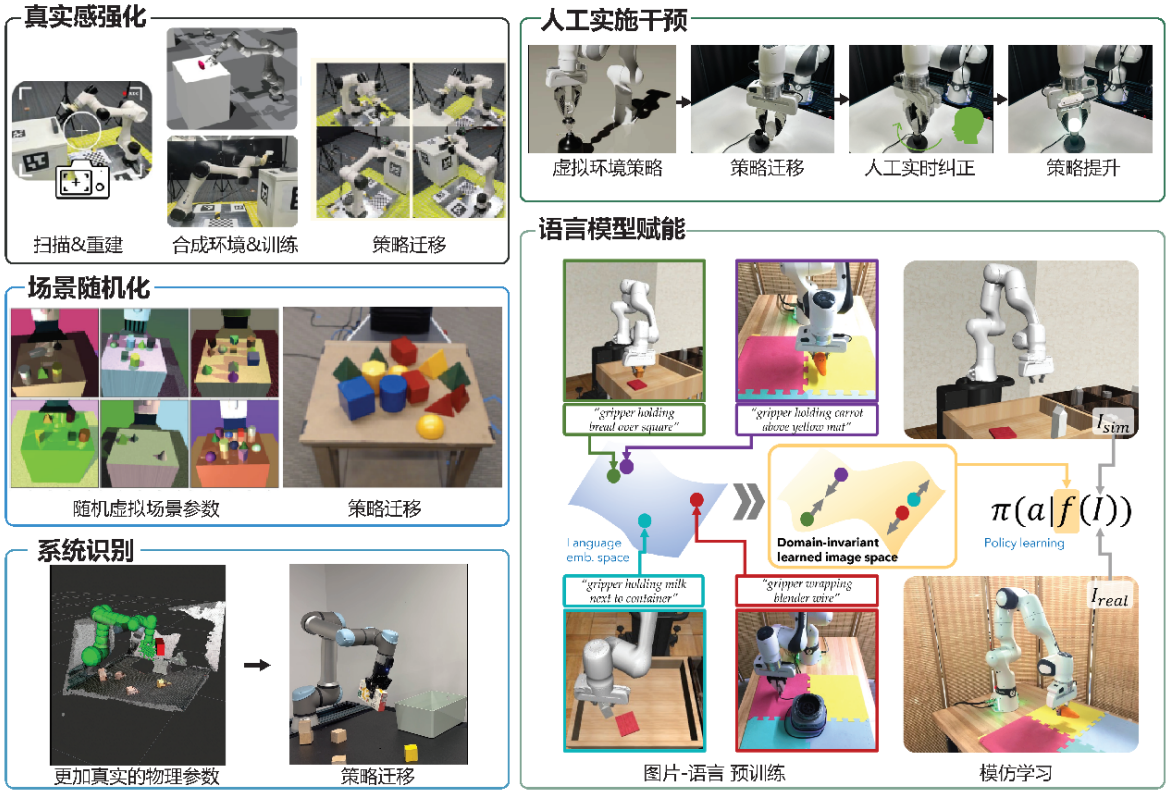

图 1-1 具身智能领域几种虚实结合的方法范例

人工实时干预[7] 是一种通过在真实场景中进行实时人工干预来纠正机器人行为，

从而缩小虚拟到现实鸿沟的方法。首先，在虚拟环境中训练以建立基本策略。随后，将这些策略部署于真实环境中，当出现错误时，人类进行实时干预和纠正行为。从这些干预中收集的数据用于训练残差策略（residual policy）。最后，将基本策略和残差策略相结合为最终策略。这种方法显著降低了对真实环境数据采集的需求，同时实现了虚拟到现实的策略迁移。

场景随机化[8–10] 通过在模拟过程中引入随机参数，增强了在模拟环境中训练的模型对现实世界场景的泛化能力。虽然虚拟和现实环境都通过相机获取视觉图像进行感知，但物体的摩擦系数和光泽度等变量使得虚拟到现实的策略迁移存在困难。因此，场景随机化方法通过在模拟训练中随机化参数，可以增强策略的泛化性，从而应对真实场景中的各种变化。

系统识别[11–13] 旨在构建真实环境的精确数学模型，涵盖动力学特性与视觉渲染等相关参数。其目的是使模拟环境与现实世界场景尽可能相似，从而让在虚拟场景中训练得到的策略可以顺利过渡到真实环境。

语言模型赋能[14–16] 用自然语言作为桥梁，通过使用图像的文本描述作为跨领域的统一信号，帮助模型学习到不受领域影响的图像特征，从而提升在模拟和真实环境中的泛化能力。首先用带有跨领域语言描述的图像数据训练一个编码器，以学习通用的图像特征。然后利用这些学到的通用特征，训练一个多领域、多任务的行为模仿策略，这个策略会根据语言指令来执行任务。这类方法利用了大量容易获取的模拟数据来弥补真实场景数据的不足，从而更好地实现从虚拟到真实环境的迁移。

第二章　具身智能的关键技术

具身智能作为人工智能领域的前沿方向，其关键技术涵盖物体操作、环境感知、任务理解与决策推理这四大核心部分，它们共同构成了机器人的“手 - 眼 - 脑”，协同支撑起智能体在现实场景中的自主行动能力。

与传统机器人存在显著差异，具身智能的物体操作有着极高要求。传统机器人的操作往往局限于特定、结构化环境下较为单一、重复的动作，而具身智能中的物体操作追求的是在复杂、动态且非结构化的真实世界场景中，能够灵活、精准地与各类物体进行交互。例如，在家庭服务场景里，具身智能机器人需要拿起不同形状、材质、重量的餐具，完成摆放餐桌、收拾餐具等一系列任务，这就要求其具备精细的力量控制与灵巧的动作规划能力。具身操作堪称当今具身智能区别于过去的关键所在，是其最核心的技术环节之一。通过先进的机械设计与控制算法，机器人的“手”能够模拟人类手部的丰富动作，实现诸如抓、握、捏、拧等多种复杂操作，从而适应多样化的任务需求。

具身感知，从范畴上属于计算机视觉的一部分，但又有着独特的侧重点。它更为关注与机器人任务紧密相关的感知信息。在复杂环境中，机器人并非需要感知所有的视觉元素，而是聚焦于对完成任务有价值的部分。以物流仓储场景为例，机器人在搬运货物时，其具身感知系统主要关注货物的位置、形状、尺寸以及周围可能存在的障碍物等信息。为达成这一目标，除了运用传统的视觉传感器，还会融合诸如激光雷达、超声波传感器等多种类型的传感器，以获取更全面、准确的环境信息，为后续的决策与行动提供坚实的数据基础。

在任务理解与决策推理方面，具身智能面临着诸多挑战。它需要对复杂长程任务进行深度理解，并自主将其拆分为一系列可执行的子任务。例如，在执行一场大型活动的场地布置任务时，机器人要理解整个活动的流程与需求，将任务拆解为搬运桌椅、布置舞台、悬挂装饰等子任务，还要合理规划执行顺序与资源分配。同时，具身智能体还需具备类人的反思与调整能力。在任务执行过程中，如果遇到突发状况，如搬运的物品过重导致移动困难，机器人应能及时反思当前策略，调整搬运方式，如寻找辅助工具或改变搬运路径等，以确保任务能够顺利完成。

在本章节中，我们将深入剖析这些关键技术，层层揭示它们如何相互协作。物体

操作依赖具身感知获取的精准信息来规划动作，任务理解与决策推理为物体操作和具身感知提供目标与方向指引。它们彼此交织、相互促进，为具身智能的蓬勃发展注入源源不断的动力，推动其从理论研究迈向广泛的实际应用。

2.1 具身感知

感知系统是生物体实现智能行为的逻辑起点，而在具身智能语境下，感知不再是孤立的信息接收，而是一个深嵌于动作-感知闭环中的动态过程，旨在构建对物理世界几何、语义及时间维度的深层动态表征。然而，这种从静态到闭环的范式转变，使得感知系统必须直面真实物理世界带来的挑战：由于单点观测的局限性，智能体必须具备主动感知与探索能力，通过改变位姿主动增强自身的感知能力；面对物理环境的复杂干扰，系统必须通过多模态信息融合利用跨模态互补性来增强鲁棒性；针对现实环境的随时间缓慢变化的特性，感知算法必须实现高效的动态环境自适应能力；此外，受限于移动机器人有限的板载算力，模型轻量化也成为了感知算法落地的重要环节。

主动感知与探索是具身智能实现自主性和适应性的核心能力。与传统的被动感知不同，具身感知强调智能体的主动性，即智能体能够根据任务需求，主动调整自身姿态、视角或交互方式，以获取更丰富、更相关的环境信息，从而更好地完成任务。主动感知系统并非全盘接收所有感知信息，而是基于任务目标和环境状态，有选择地关注特定信息[17]。主动感知强调感知过程与动作的紧密结合。智能体通过控制自身的视角和交互方式，主动地获取有利于理解环境的感知信息。例如，机器人可以通过调整头部姿态来获取更清晰的图像[18]、通过触摸来识别物体的属性[19]、或者主动调整相机以追踪操作物体[20]。除了短期内辅助感知的主动行为，主动感知还包括更长时期内的规划探索。例如 MP5 算法[21] 在 Minecraft 中实现了精细的环境感知策略规划，开发了目标导向的主动感知算法，能够处理更复杂的任务。ActiveGAMER[22] 利用 3D 高斯溅射的渲染特性评估环境重建任务的信息增益，驱动机器人主动规划最佳视角，实现了更优的全局重建性能。在主动感知与智能体行为的端到端学习方面，APPLE[23] 设计了一个主动感知的策略学习方法，并验证了其在多种任务中的可靠性。

在具身智能体系中，多模态感知正在从早期的传感器堆叠向更深度的全感官表征耦合转变。单一模态的局限性要求系统必须整合视觉语义、激光雷达、触觉反馈及本体感受等多源数据，以实现准确的环境评估、意图理解及安全决策[24,25]。为应对数

据异构性与关联建模挑战，早期研究主要侧重于特征级联[26]和决策级融合[27]，以及利用交叉注意力机制实现异构特征对齐的统一空间表征[28]。近年来，具身感知更强调在非结构化交互中的感知鲁棒性与多传感器的协同对齐。在感知能力的拓展方面，非视域成像技术[29]通过三维模糊核建模与空间相关性重采样，实现了对遮挡目标的高分辨率实时重建，为机器人在视觉受限场景下的感知提供了新的可能性。在传感器协同方面，Metasensor[30]提出了一种基于传感器演进的架构，通过自适应调度机制使智能体能在视觉受限时无缝切换至其他感知分支，实现了面向任务的最优感知切换。同时，VTDexManip[31]与视觉-触觉预训练框架[32]的研究表明，通过视触觉几何深度对齐，可为近距离精细操纵提供亚毫米级的反馈，有效弥补传统视觉检测在遮挡环境下的感知盲区。这些进展标志着多模态感知已跨越简单的语义感知任务，迈向具备持续时空一致性的全感官物理认知。

在动态环境中，感知模块的自适应能力至关重要，这使得智能体能够实时调整其感知策略以应对环境的变化。这种适应性主要通过在线学习和迁移学习来实现。在线学习使智能体能够通过与环境的持续交互来逐步优化其感知模型。通过持续环境交互，智能体可以持续提升感知能力，并学会更有效地利用信息来完成任务[33]。实现这一目标需要解决两个关键问题：首先是高效收集信息丰富的训练样本，常用的标准包括语义分布和不确定性[34]；其次是如何利用这些样本来更新感知模型，现有研究探索了逆动力学预测目标[35]、时空一致性约束[36]、探索行为学习[37]等方法。最近的研究[38]则开始尝试完善整个自进化具身智能框架，提出了记忆自我更新、任务自我切换、环境自我预测、具身自我适应和模型自我进化这五个重要组成部分，展现了具身智能进化的潜力。

关于具身智能感知的轻量化研究，其正在从早期的静态模型压缩迈向深度耦合现实任务与硬件底层特性的全方面优化，核心目标是在边缘侧有限的功耗与算力约束下，维持高频、高精度的环境感知建模。在算法层面，针对视觉 Transformer 计算复杂度随词元数量呈二次方增长的瓶颈，研究界引入了层级化提前退出机制与跨模态引导裁剪等任务驱动的加速技术[39,40]。这些技术能够模仿人类视觉从粗粒度到细粒度的处理过程，依据环境语义难度及实时语言指令动态剔除比例显著的冗余视觉词元，确保计算资源精准聚焦于目标交互区域，从而在维持性能的前提下大幅降低显存与带宽压力。此外，轻量化设计也进一步下沉至硬件底层[41–43]，针对硬件特性实现有针对性的加速设计。这种从算法逻辑到芯片架构的深度协同，缓解了模型在资源受限终端部署的实时性瓶颈。

2.2 具身推理

具身推理是指具身智能体通过多模态感知（如视觉、触觉、本体觉等）实时捕捉环境状态，融合行为目标与历史经验，从而将复杂任务分解为可执行的任务规划、并进一步类人地动态修正执行错误的认知框架。其典型应用涵盖机器人自主导航、操作顺序控制、人机交互等多个方面。通过闭环反馈机制，智能体在执行动作后持续监测和评估实际执行结果，并利用这些反馈信息更新内部认知模型和决策策略，实现自我反思调整和优化。这使得智能体能够自动分解复杂任务、适应环境变化并不断从经验中学习。例如，一个服务机器人在家庭环境中执行“准备早餐”这一复杂指令时，不仅需要理解自然语言描述的目标并将其转化为取物、烹饪、摆盘等一系列具体操作，还需根据实时感知到的环境状态调整计划。在这一过程中，具身推理将认知决策与物理反馈紧密耦合，使机器人在真实世界中表现出更高的鲁棒性和适应性。早期的研究主要基于符号推理和规则系统，依赖人工预先定义的规则、状态机或经典 AI 规划器（如 PDDL 规划），这种方法往往难以应对开放环境、抽象目标和非结构化环境。

近年来，随着大语言模型和多模态大模型的推理能力不断增强，推理能力不仅在自然语言处理领域取得了显著成果，也展现出在机器人规划与交互中的巨大潜力。研究人员正探索如何将大模型的强大推理能力与机器人的感知和控制相融合，使机器人能够理解高层指令并推理出具体的行动序列，从而更有效地完成任务。借助大模型在知识获取和推理方面的优势，机器人可以在零样本或少样本条件下理解复杂指令并自动生成新任务的执行步骤，从而大幅减少人工设计预定义规则的工作量。这种从符号人工编排向数据驱动智能规划的转变，使具身推理框架具备了更高的通用性和适应性。

大模型在具身推理中的作用主要体现在三个方面：

（1）语义理解与目标分析：将模糊的自然语言需求转化为明确的目标是推理的基本表现形式。大模型通过对海量自然语言语料的学习，能够准确捕捉用户指令中的隐含需求，将模糊的描述转化为清晰、量化的目标。这种能力使得智能体在接收到“我渴了”这样的抽象指令时，能够转化为“倒一杯水并送到人的面前”这样的具体而清晰的目标，从而能够更好地呈现人和具身智能体之间的人机交互智能。

（2）原子动作分解：将总体任务目标分解为可逐步推进的子任务，并转换为机器人低层技能可执行的原子动作序列，是推理能力的另一个重要体现。大模型需要将整体目标拆解为一系列连贯、可操作步骤。例如，谷歌提出的 SayCan[44] 方法，对于用户

提供的高层自然语言指令，大模型会输出一系列可能的子任务描述，然后机器人依据已学习的技能库和环境状态，通过机器人学中的价值函数或可行性模型（affordance）对这些候选子任务进行打分，来选出当前最合适的原子动作。例如，面对指令“我把饮料打翻了，你能帮我吗?”，大模型能够规划出连贯的行动序列，诸如“走到柜台，拿起海绵，擦拭液体”等步骤。在每一步执行后，大模型根据新的环境反馈继续规划下一步，直到任务完成。这种将大模型的知识与推理能力同机器人对物理世界的执行能力相结合的方式，使机器人可以完成诸如“收拾洒出的饮料”这类复杂任务，并显著减少了不切实际或不可执行的错误。

（3）反思与调整：根据子任务的执行情况和实时环境反馈，调整原子动作规划，优化任务执行流程，则是具身推理呈现给具身智能体的环境感知的核心能力。例如，谷歌提出的 ReAct[45] 方法利用大模型生成交错的推理思维链和原子动作，使模型能执行动态推理来创建和调整行动计划，同时与外部环境交互获取额外信息，在具身问答和互动决策等任务中取得了显著的推理性能提升。斯坦福提出的 Text2Motion[46] 方法，将大模型与物理可行性预测模块相结合，在生成动作序列时既确保语义逻辑正确，又验证现实操作的可行性。系统生成多条路径，通过评估每条路径在实际环境中的执行效果并动态调整具体步骤，从而保证机器人既满足指令要求，又能克服真实环境中的物理限制。谷歌提出的 VLP[47] 方法进一步结合视觉信息，并利用视频生成基础模型在任务执行前预演行动过程，通过树搜索评估每一步的合理性与可行性，使机器人在实际执行前便能识别潜在的风险，从而调整任务规划。在失败检查与反思方面，REFLECT 框架通过 LLM 分析视觉基础模型反馈的检测结果，检查子目标是否达成，进而利用大语言模型进行失败原因推理和计划的实时纠正，该方法不仅能够检测执行错误（如物体被意外掉落），还能识别规划错误（如选择了错误的目标物体），为具身推理系统提供了重要的自我纠错能力。

然而，上述方法普遍依赖预定义的动作技能库，这在开放场景中存在显著的局限性，很难穷举所有潜在动作技能。为突破这一局限，研究者们提出了代码生成的范式，直接从自然语言指令生成机器人执行策略代码，从而减轻对预设动作库的依赖。Code-as-Policies[48] 作为典型代表，利用大语言模型编写基于自然语言的控制代码，动态调用机器人 API，生成适应性行为策略。RoboCodeX[49] 方法进一步结合了视觉反馈信息，构建了多模态代码生成框架，将高级人类指令与视觉观测结合，分解成以对象为中心的操作单元，再转化为适用于不同机器人平台的控制代码。代码生成技术将大模型的语义理解能力与机器人执行能力有机融合，实现了从自然语言到可

执行策略的高效转化，使机器人在复杂环境中展现出前所未有的任务适应性和执行灵活性。

新一代具身推理系统正朝着多模态深度融合和自适应策略方向不断迈进。在 VoxPoser[50]、OmniManip[51] 和 ReKep[52] 等最新研究中，研究人员利用视觉、语言等多源信息构建了实时、闭环的决策模型。VoxPoser 通过生成三维价值图，将环境中的可操作区域与障碍物直观映射，使机器人在任务规划时无需依赖预设动作模板，而是基于当前场景直接计算出最优轨迹；与此同时，OmniManip 和 ReKep 则利用关键点约束将复杂任务拆解为一系列精细化的动作子目标，并通过实时视觉反馈不断优化运动路径，从而在动态环境中实现更加精准的控制。依托大语言模型和视觉模型的强大推理能力，这一全新的范式不仅具备零样本任务适应性，还能够通过闭环反馈机制及时感知偏差，根据实际执行情况动态调整操作细节。系统能够将模糊的自然语言指令转化为明确、量化的目标，并进一步将整体任务分解为一系列可逐步执行的原子动作。通过直接生成可执行策略代码，这种方法大大降低了对预定义动作库的依赖，为机器人在开放场景下自主决策提供了全新的思路和途径。然而，通用多模态大模型在直接执行具身推理任务时，往往只依赖预训练期间获得的知识，缺乏专门的具身任务规划能力，因此需要额外的反馈模块（如人工提示）来修正不合理的规划。更直接的方法是利用视觉语言数据对预训练模型进行微调，使其更适应真实环境下的任务需求，从而打造更强大的具身大模型。例如，谷歌提出的 Palm-E[53] 将状态、图像和文本统一编码，通过自注意力机制生成任务规划；而 EmbodiedGPT[54] 则依托 EGO4D[55] 数据集，构建了包含视频片段与原子操作语义配对的 EGOCOT[53] 具身思维链数据集，经微调后其规划性能显著提升。在视觉语言大模型的基础上，最近的研究还引入了 RT-1[56]、RT-2[57]、RT-X[58] 和 $\pi 0$[58] 等视觉-语言-动作大模型。它们通过利用互联网上丰富的视觉推理任务数据进行预训练，并结合端到端输出底层动作的策略优化，实现了从高层指令到具体动作的高效转化。后续章节将详细展开具身大模型在各类任务中的表现与应用实例，进一步探讨其在具身推理和多模态交互中的前沿进展。

2.3 具身操作

具身操作是新一代智能机器人发展的核心技术，是机器人能否完成更多新任务的关键，代表了具身智能的重要发展方向。当下，除了上述 VoxPoser[50]、OmniManip[51] 和 ReKep[52] 等非端到端的研究，每一时刻逐步输出动作的端到端操作模型是下一

步技术发展的主流共识，其中视觉-语言-动作大模型（Vision-Language-Action Model，VLA）（图2-1）通过融合实时视觉感知、自然语言理解和动作控制三大模块，为多样化机器人任务提供了通用解决方案，被视为新一代机器人智能中枢，在全球学术界和产业界引发高度关注。在 2026 年，随着视频模型的进一步成熟，将视频预测和动作预测融合到同一个模型中的世界动作模型（World-Action Model，WAM），提升了基于仅动作预测的 VLA 的泛化能力，成为了业界关注的新热点。

当前 VLA 模型主要包括三大技术路线：

（1）基于 VLM + 动作模型架构

采用视觉-语言大模型（Visual-Language Model, VLM）解析任务指令和视觉信息，通过知识推理引导动作模型生成控制指令。典型应用如叠衣服、餐盒整理等精细化操作任务。

（2）基于 VGM + 动作模型架构

基于视频生成模型（Video-Generation Model, VGM）预测任务执行过程的视觉轨迹，通过视频表征逆向推导控制指令。适用于需要预判连续动作序列的复杂操作场景。

（3）基于 VLM+Latent+Action 架构

为了充分利用各类异构数据，增强策略的泛化能力，将互联网图文数据和机器人数据混合训练，通过通过预测隐式动作标记（Latent Action Tokens) 来弥合图像-文本输入与机器人执行动作之间的鸿沟。

三种技术路线在 2024-2025 年间取得突破性进展，但目前尚未形成统一范式，呈现多元化创新格局。

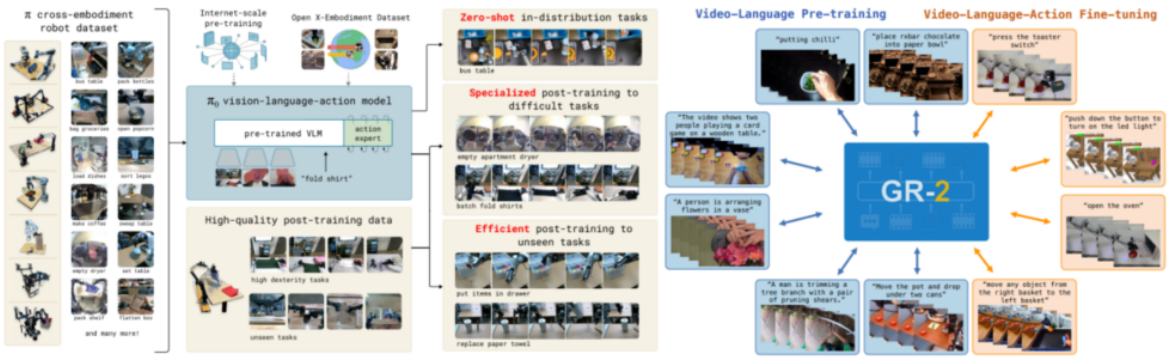

图 2-1　Physical Intelligence 的 π0 模型以及字节跳动的 GR-2 模型

基于 VLM + 动作模型架构的技术路线：2024 年中，Physical Intelligence 发布了 π0 模型，依托互联网级数据训练的 VLM 主干网络，通过跨构型预训练 + 单构型后训练，实现洗衣机取衣、餐盒折叠等家庭服务能力。清华大学发布了 RDT 模型，以动作模型为核心架构，通过跨构型数据集从零开始的预训练，结合双臂机器人后训练，成功完成双臂倒水、游戏手柄操控等任务。2025 年，Figure 发布了 VLA 部署到人形机器人作业，该模型采用 VLM 与动作模型的解耦设计，感知数据并行输入这两个模型。该架构完整部署两个模型能够支持杂乱桌面整理等复杂任务，也可通过只部署动作模型实现包裹分拣等基础任务。

基于 VGM + 动作模型架构的技术路线：与基于 VLM+ 动作模型的架构类似，基于 VGM + 动作架构的技术路线也有两类。单一模型方面，字节跳动发布了 GR-2 模型，采用大规模人类视频数据预训练 + 小规模机器人数据后训练的方式，实现可泛化的物品分拣任务；在分层架构方面，学术界产生了一些相对较小但也极具代表性的模型，例如清华大学提出的 ATM，使用关键点光流作为中间表征解耦视频预测和动作模型，实现少样本机器人数据微调下的技能学习；南洋理工大学提出的 FLIP，则使用稠密光流作为中间表征引导动作生成。

基于 VLM+Latent+Action 的技术路线：为了充分利用各类异构数据，增强策略的泛化能力，2025 年初智元提出了 Vision-Language-Latent-Action (ViLLA) 具身操作大模型。与 VLA 架构相比，ViLLA 通过预测隐式动作标记（Latent Action Tokens），弥合图像-文本输入与机器人执行动作之间的鸿沟。在真实世界的灵巧操作和长时任务方面表现卓越，远远超过了已有的开源 SOTA 模型。ViLLA 架构是由 VLM+ 混合专家（MoE）组成，其中 VLM 借助海量互联网图文数据获得通用场景感知和语言理解能力，MoE 中的隐式规划器（Latent Planner）借助大量跨本体和人类操作数据获得通用的动作理解能力，MoE 中的动作专家借助百万真机数据获得精细的动作执行能力。在推理时，VLM、Latent Planner 和 Action Expert 三者协同工作。

相比于 VLA 模型仅生成动作，WAM 模型的目标是对动作-下一个状态形成的分布进行建模。因为下一个状态通常是图像，所以采用视频模型作为主干，并利用注意力机制实现动作分支基于视频生成过程中的特征对动作进行预测。这条技术路线相比于 VLA 模型能够接受来自下一个状态预测的梯度，数据利用更高效。Nvidia 遵循这条路线，推出了 DreamZero，实现对于新任务，仅需 10-20 分钟的演示数据即可带来性能提升。蚂蚁灵波科技的 LingBot-VA 也是 WAM 路线的代表性工作，通过因果注意力的设计，实现推理的加速，性能上获得了类似 DreamZero 的结论。

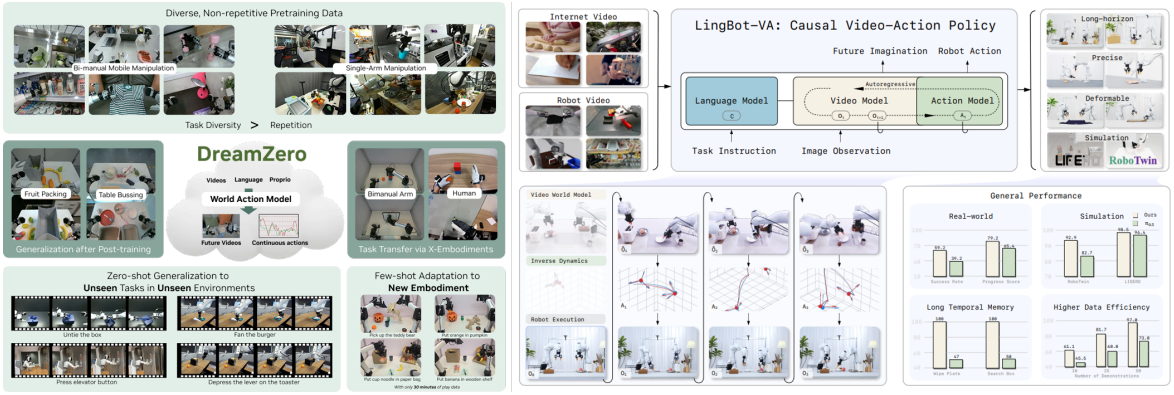

图 2-2　NVIDIA 的 DreamZero 和蚂蚁灵波的 LingBot-VA

2.4　具身导航

具身导航是指在动态环境中智能体感知时空关联内容并与环境交互，进而移动到由特定坐标[59]、物体[60]、语言描述[61] 或图像[62] 指定的目标位置。该任务中，智能体处于未知环境中的某一随机初始位置，需要根据先前的导航经验来导航到指定的目标方位。具身导航的研究问题集中于如何学习导航的先验知识，这种先验知识应当在未知的、多样的环境中具有良好的泛化性，从而来帮助智能体基于这些经验来快速、高效地导航到目标位置。

与传统移动机器人中的几何导航模块相比，具身导航更加关注“引导机器人本体前往一个能够完成任务的位置”。传统导航通常以二维栅格或高精度地图为基础，在已知或弱变化环境中规划一条无碰撞、几何意义上的最短路径，导航目标多为预先给定的坐标点或静态路标，对机器人本体形态、相机视角、可操作区域以及任务语义的考虑有限。具身导航则将导航放在具身任务整体闭环中：一方面，目标可能以“找到某个可见/不可见物体”、“根据一句自然语言描述抵达某区域”等高层语义形式给出，智能体必须联合视觉、语言与本体状态进行推理；另一方面，导航结果不仅要求机器人抵达某个坐标，还要保证后续操作（如抓取、开门、插拔、电梯交互等）的可达性和舒适性，例如需要面向桌面、预留机械臂运动空间、兼顾视野覆盖和避障约束。从系统架构看，简单地在具身机器人上“直接安装传统导航模块”这种方式往往将导航视作一个与感知、操作割裂的黑盒子：上层只给出目标坐标，导航模块仅返回轨迹或速度控制。这种方式在结构化、规则化场景下即可满足需求，但在开放环境中存在明显局限：其一，缺乏对语义目标和任务约束的显式建模，难以处理“按语义指令去往某类物体附近”、“为了后续抓取而调整姿态”等具身任务；其二，缺乏与操作控制的

一体化协同，导航结束后往往还需要额外的位姿微调和复杂的状态机衔接；其三，传统模块多依赖手工建模和调参，在动态人群、强遮挡、传感器噪声等场景下面临鲁棒性瓶颈。相较之下，具身导航通常通过端到端或模块化学习，将视觉-语言表征、时空记忆和动作策略统一到一个共享表征空间中，在同一套框架下同时优化“去哪、怎么去、到达后如何继续执行任务”等目标，在未知环境、长时序、多目标任务中展现出更好的泛化能力和任务完成效率。

为了学习可泛化的导航经验，当前的方法可以分为需要训练和免训练两种类型，其中需要训练的方法可进一步划分成基于端到端强化学习和基于模块化学习两种类型。端到端强化学习方法将导航经验建模成端到端的策略网络，并通过增加情境记忆[63] 和改进视觉表示[64] 等方式来增强策略网络，整个策略网络通过在训练环境中的端到端强化训练方法进行学习。然而，端到端方法的能力受制于强化学习的采样低效、训练成本高等局限性。此外，由于缺乏显式的记忆地图，这还限制了其在复杂环境和长时序导航下的泛化能力。另一类是基于模块化的方法，这种方法通过构建语义地图[60] 来精确记忆空间关系，然后通过监督学习预测语义图边界[60] 或者通过自监督学习预测语义地图的未知区域[65]，据此构建导航点预测模块并进行路径规划，从而寻找最有价值的导航方向以实现高效的探索。模块化的方法虽然通过引入记忆模块增强了泛化性，但面向新目标仍需要进行额外训练，所以泛化性提升有限。免训练的方法不需要额外的训练且可以适用于开放集合的目标，这些方法借助于视觉-语言模型（VLM）或大语言模型（LLM）来获取导航经验并实现动作的预测。在这类工作中，大量的工作将所感知的视觉内容转换为文本，然后将文本形式的环境内容以及历史信息输入大语言模型来预测导航动作。具体而言，基于 VLM 的方法[66] 利用视觉-语言预训练模型[67] 计算实时观测与目标语义之间的相关性分数，从而引导智能体向分数较高的区域导航。基于 LLM 的方法[68] 利用 LLM 强大的先验知识，并通过提示每个边界附近的物体类别来估计接近目标的概率，引导智能体不断接近高概率边界，进而不断接近目标。免训练的方法是灵活的，模型结构上也有着较为统一的框架，且在真实环境的应用具有更广阔的前景，但其导航性能低于训练方法。

为进一步缩小模拟环境训练的模型迁移到真实环境的泛化差距，并提升真实环境中的导航性能，依赖于导航多模态数据来端到端微调大语言模型的方法得到了大量的关注。这类方法将导航轨迹视频信息输入视频表征模型，以获取结构化顺序信息，并联合语言指令等多模态信息共同输入大语言模型，进而端到端训练视觉语言行为模型，完成动作预测。面向真实环境下的具身智能系统，当前具身导航的研究

方向从关注在未知模拟器环境下有限集合的导航能力，逐渐延伸到关注于开放导航、真实环境下的导航、以及导航协作等更复杂、综合、泛化、通用的导航情景中。

2.5 强化学习

近年来，强化学习（RL）在具身智能领域取得了显著进展。具身智能的核心理念在于智能体通过与物理环境的交互，学习并完成任务，而强化学习作为一种基于环境交互、通过试错和奖励机制进行优化的学习范式，已经成为实现具身智能的核心技术之一。随着深度学习、计算能力和算法优化的突破，强化学习技术从实验室研究走向了实际应用，推动了机器人在自主决策和环境适应方面的发展，不仅为多个具身智能任务（包括导航、操作、运动、交互）提供了技术支持，也为未来的创新奠定了坚实基础。

导航任务要求智能体在复杂环境中自主规划路径并避开障碍物。传统算法通常依赖视觉输入，机器人通过图像数据感知环境并做出决策。强化学习则将导航任务建模为马尔可夫决策过程，通过与环境的交互帮助智能体学习最优路径规划策略。例如，UC Berkeley 提出的 NoMaD 框架基于扩散模型进行策略学习，同时处理目标导向导航与无目标探索，使机器人在未知环境中能灵活搜索并达到指定目标[69]。Meta 的 NWM 框架则通过结合图像扩散生成模型，增强生成式强化学习策略，从而实现机器人在更开放场景中的导航规划[70]。

在操作任务中，强化学习使机器人能够高精度地完成物体抓取、搬运和堆叠等任务，并在复杂环境中逐步优化操作策略。近年来，将 VLA（Vision-Language-Action）模型与强化学习相结合的方法能够整合视觉信息、语言指令与动作决策，令机器人不仅具备物体识别能力，还能理解并执行基于语言的任务指令，从而提升操作效率。例如，DeepMind 的 RoboCat 通过预训练模型生成数据促进后续训练，并结合强化学习的自适应特性，形成自我改进的循环，大大降低了对大量标注数据的依赖[71]。Stanford 提出的 HumanPlus，将模拟环境中的强化学习与现实世界对真人行为的模仿学习结合，实现了具身智能的自主技能学习[72]。

强化学习在运动控制任务中已广泛应用于提高机器人在动态环境中的稳定性与灵活性，尤其是在腿足机器人和飞行控制领域[73,74]。借助强化学习的反馈机制，机器人能够通过与环境的高频互动逐步优化控制策略，从而实现高效的运动规划和精确控制[75]。四足机器人领域，UC Berkeley 提出 AMP 算法，使用模仿对抗网络替代传

统强化学习算法中复杂的奖励设计，进一步提高了强化学习方法在运控任务中的效率和可拓展性[76]。上海交通大学提出的 HugWBC 框架通过统一的整体控制策略，实现了人形机器人多种自然步态的精细控制，并支持上半身外部控制的实时干预，从而提升了机器人运动的灵活性和鲁棒性[77]。

在交互任务中，强化学习的应用使得智能体能够与人类或其他智能体进行更加有效的合作与交流。通过模拟与反馈机制，强化学习不断优化智能体的行为策略，使其在交互中更加自然、智能且高效。工业环境中，Amazon Robotics 的仓库系统使用多个机器人协同工作，完成物品搬运、存取和分拣任务，其中强化学习帮助这些机器人实现了任务分配、路径规划和实时适应环境变化等功能，从而大大提升了仓库操作的自动化水平和效率[78]。在人机交互领域，RHINO 框架通过层次化学习，使得人形机器人能够实时反应，并根据人类的指令和行为动态调整任务执行，为未来灵活和安全的人机交互奠定了基础[79]。

2.6 具身交互

早在 2001 年，在具身智能概念尚未形成之前，具身交互（Embodied Interaction）理论便由加州大学欧文分校的保罗 · 杜里什（Paul Dourish）教授提出，其代表性著作《行动在何处：具身交互的基础》[80] 系统阐述了具身认知作为人机交互设计理论基础的重要意义，对过去十余年的交互设计理论与实践产生了深远影响。在该理论框架下，认知被视为源于身体、环境与行动之间的动态耦合过程，而非孤立的符号计算。随着具身智能的快速发展，具身交互的内涵被进一步拓展与深化，不仅涵盖具身主体与物理环境之间的交互过程，也包括人类参与下的人机在环（Human-in-the-loop）交互模式，从而形成融合环境适应与人机协同的统一交互范式。

具身主体与环境的交互是具身智能框架中的核心环节之一。智能体通过多模态传感器对环境进行细粒度感知，并基于感知结果在强化学习框架下进行决策与动作选择。在这一过程中，智能体与环境的每一次交互都会产生反馈信号，该反馈本质上是对其行为结果的评估。在强化学习语境下，这类反馈通常被形式化为数值奖励或惩罚，用于驱动策略的持续优化。通过“感知—决策—行动—反馈”的闭环机制，具身智能体能够在复杂动态环境中不断调整行为策略，实现对环境的自适应与持续学习。例如，在具身抓取任务中，是否成功抓取目标物体构成交互反馈；在具身导航任务中，是否到达目标位置或发生碰撞同样反映了交互结果。

在具身对话（Embodied Dialogue）任务中，智能体不仅需要与环境进行紧密交互，还需通过自然语言与人类用户进行沟通，以实现更加自然与高效的人机协作。具身对话强调智能体在理解并执行人类指令的同时，具备主动交互能力，例如在指令存在歧义时进行澄清，或在信息不足时主动请求补充，从而提升任务执行的准确性与鲁棒性[81]。该任务通常要求智能体具备多模态感知与融合能力，能够综合利用环境信息、语言输入以及人类反馈进行决策。从交互机制上看，具身对话关注人机之间的双向、多轮协同过程。有效的交互依赖于动态演化的对话机制[82]，即智能体在任务执行过程中能够主动发起询问，并根据用户反馈不断修正行为策略。这种双向交互模式突破了传统单向指令执行的范式，使智能体逐步具备协同决策能力。为推动该方向的发展，研究者构建了一系列支持双向对话的基准任务与数据集，如 DialFRED[81]。进一步地，一些研究开始模拟真实家庭环境中的人机交互场景，不仅包含人机对话，还引入了大量人类之间的协作对话[83]，从而显著提升了任务复杂度与现实性。这类数据涵盖复杂指令理解、多步任务规划与动态交互过程，为具身对话研究提供了重要支撑。具身对话的典型交互形式如图 2-3所示。

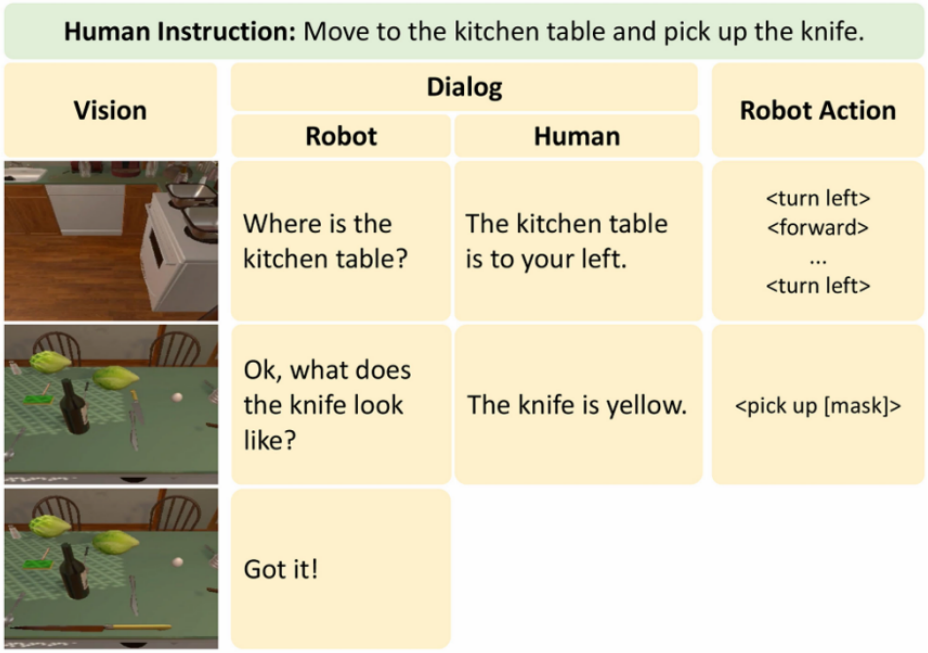

图 2-3　具身对话示例

人机在环的具身交互（Human-in-the-loop Embodied Interaction）是一种人类与智能体深度协同的交互模式，其核心在于通过引入人类参与，实现对智能体学习、决策与执行过程的有效调控与增强。该模式主要可分为两种典型形式：

一类是基于人类反馈的在环机制，即人类通过主动提供反馈、指导与监督，参与智能体的训练与决策过程，从而提升其在复杂环境中的适应性与安全性。在该模式下，人类通常不直接参与具体任务执行，而是以“外部监督者”的角色对智能体行为进行校正与优化。近年来，这一机制已广泛应用于大语言模型及具身大模型（如RT-2[84]）中，尤其在处理高不确定性与长时序决策任务时，能够显著提升模型性能与稳定性。同时，人类反馈的引入还有助于增强模型的可解释性与可控性，从而提高系统整体的可信度[85]。另一类是人类深度参与的协同交互模式，即人类直接进入具身主体的工作空间，与智能体共同完成任务。在该模式下，人类与智能体分别发挥各自优势：智能体依托其感知与计算能力执行具体操作，而人类则在复杂决策、异常处理及高风险环节提供关键支持。这种协同方式通常与具体应用场景高度耦合，例如在医疗手术、工业协作等高精度任务中具有重要应用价值。以手术机器人为例，具身交互不仅涉及机器人与物理环境的交互，还包括与人类操作员（如外科医生）的紧密协同。Long 等人[86]（图2-4）构建了一个支持高质量人机交互的仿真平台，将人类因素显式引入具身智能系统中。实验结果表明，该人机在环交互模式不仅能够提升手术机器人的自主性，还能够通过人类在关键决策与复杂操作中的参与，有效保障系统的安全性与可靠性。总体而言，人机在环的具身交互通过融合人类智能与机器智能，在提升系统性能的同时强化了安全性与可控性，为具身智能在高风险与高复杂度场景中的落地应用提供了重要支撑。

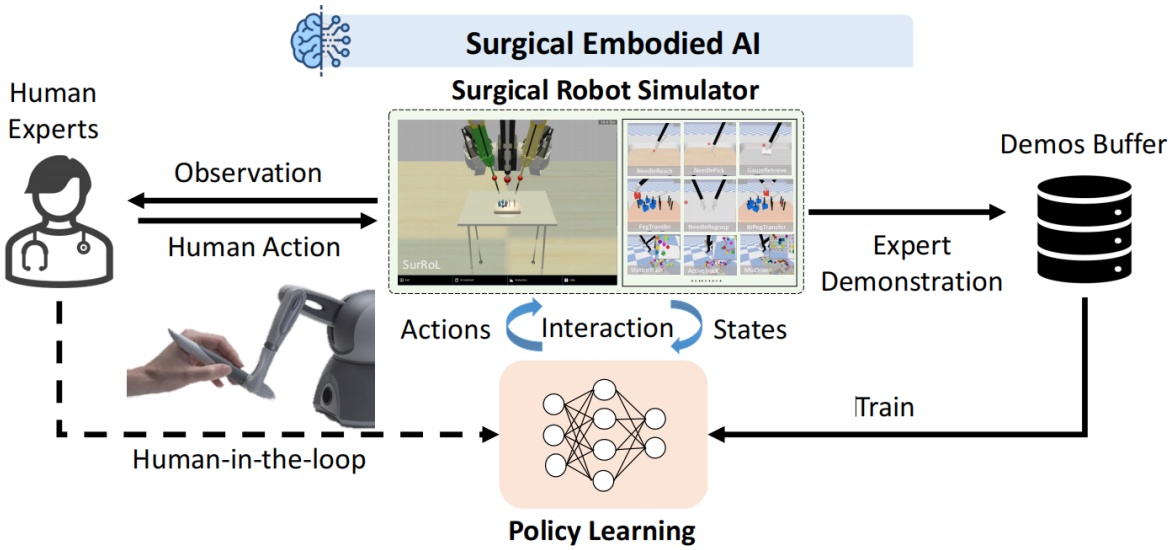

图 2-4　人机在环的手术场景具身交互流程

2.7 群体具身智能

群体具身智能是机器人集群技术与具身智能深度融合的前沿方向，凭借多机器人分工协作的核心优势，已成为自动化解决复杂工程问题的关键技术手段，可在复杂任务场景中展现出超越单一机器人的作业效率与任务完成度[87]。近年来，机器人集群系统凭借广阔的应用前景成为学界研究热点，浙江大学团队研发的四旋翼无人机集群系统实现了密集野外环境下的类鸟群自主导航，可在树林间灵活穿梭[88]；香港中文大学（深圳）团队提出的模块化自重构机器人，能够通过多机连接组合适配复杂非结构化环境[89]。机器人集群智能研究正从传统分布式、集中式协同规划与控制的协同运动层级，向协同认知、自主决策与复杂系统调度的协同认知与作业阶段跨越。

为进一步提升集群系统认知、协作与决策能力，研究人员借助大模型理解与推理能力，开展基于集群基座模型的智能决策架构研究，为集群智能注入“认知与推理引擎”。如图 2-5所示，群体具身智能平台主要由大脑、小脑及单元平台组成，集群具身大脑负责决策，通过机器人本体信息共享、人机交互、任务与环境等多维度分析器强化智能体对物理世界与任务的认知能力；依托顶层决策大模型进行深度逻辑推演，实现集群核心推理与决策功能，并通过集中式或分布式方式进行任务调度与分配；集群小脑负责本体的执行策略，结合多模态数据与高精世界模型，构建感知-动作-操作闭环的端到端具身策略，赋予集群在动态环境下自主避障与重构能力；最终，通过集群大脑、协作小脑及平台本体感知控制能力，实现多机协同感知决策、动态环境自主避障重构、复杂任务分工协作的核心目标，构建高鲁棒性、高协同性、高适应性的具身智能集群系统，推动机器人集群技术从单一任务向复杂具身操作任务的跨越式发展。

2025 年以来，群体具身智能技术实现突破性发展，持续革新大模型驱动的集群决策架构，智源研究院提出的 RoboOS 框架通过云端大脑与端侧小脑的协同架构，解决了异构机器人集群的统一调度难题[90]；LaMMA-P 框架首次实现大语言模型与 PDDL 规划器的深度融合，大幅提升多智能体长时任务规划成功率[91]；集群具身推理能力持续跃升，Arcadia 智能闭环学习框架[92] 与 Uni-Walker 模型[93] 着力解决动态环境下集群系统的持续学习与灾难性遗忘问题；香港中文大学（深圳）研究团队提出意图对齐模仿学习方法，解决多种异构机器人（含无人机、无人船、轮足机器人、人形机器人、机械手等）意图级行为适应与协作难题[94]，为群体具身智能、异构机器人集群协作提供了全新的技术范式。

群体具身智能正经历从“协同移动”到“协同认知 + 协同作业”的深层变革，系

统从静态规划走向动态适应、从同构协同走向异构融合、从人类设计走向自我进化，未来方向包括知识图谱与大模型的深度耦合、多智能体世界模型等方面研究，为灾害救援、智慧物流、林业探测、国防科技等应用奠定技术基础。

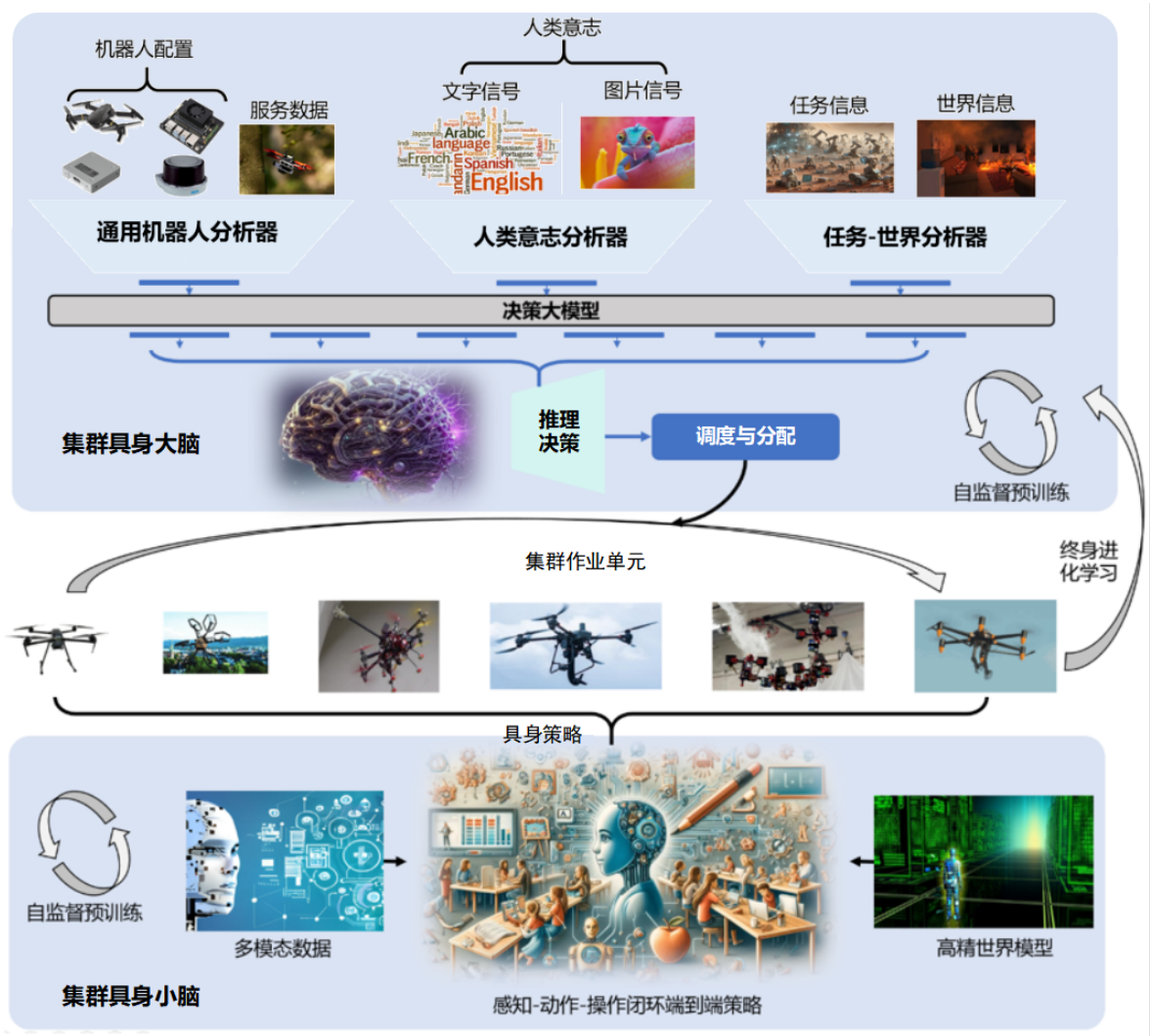

图 2-5　具身智能集群系统框架图

2.8　具身世界模型

随着大语言模型（LLM）在自然语言处理领域的快速发展[95]，以及大规模视觉模型和视觉语言模型（VLM）对计算机视觉任务的零样本/小样本泛化能力的提高[96–99]，学术界围绕如何实现通用人工智能（AGI）[100] 进行了大量的讨论。在此背景下，世界模型已经成为机器人学中的一种变革性范式，使机器人能够在复杂环境中有效地感知、预测和执行任务，其面向具身智能的自主智能整体架构如图 2-6所示。

机器人领域通过大语言模型的相关技术[44,101,102]构建了基础模型[103]，这赋予了机器人系统世界感知、任务规划甚至运动控制的能力。虽然这些模型展示了捕捉各种世界知识的能力，但仍缺乏一个内部世界模型来预测世界状态并与现实世界交互。Ha 等[104]通过建立世界模型，对现实世界进行感知和建模，从而深入了解其潜在机制；而 LeCun[100]认为，世界模型还应该具备预测未来状态的能力，以便为决策提供参考；世界模型预测[105,106]等工作直接预测世界的未来状态，使机器人能够预见可能的状态变化并主动反应，为机器人直接与现实环境进行交互和学习提供帮助。

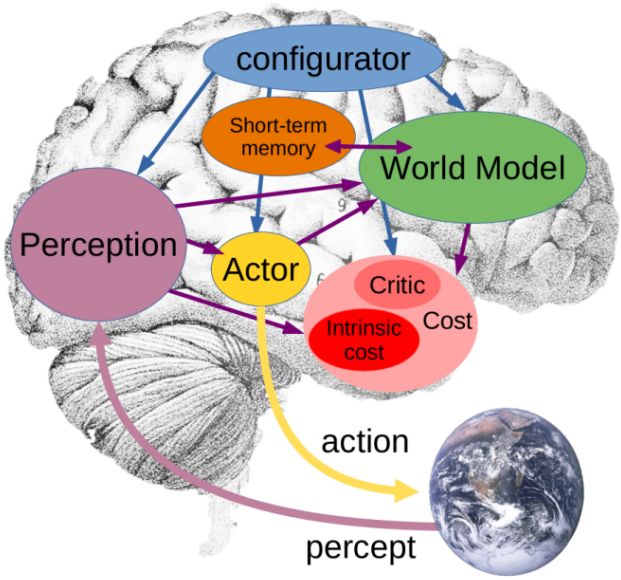

图 2-6　自主智能架构

为了帮助机器人感知和理解世界，视觉 Transformer（ViT）[107,108]等视觉模型将图像的视觉特征映射至模型空间中，使机器人能够识别环境中的关键物体；Robo Craft[109]通过图神经网络捕捉底层系统的结构；PointNet 工作[110,111]对三维点云进行函数编码以捕获环境的空间特征；Gornet 等[112]的工作将通过局部搜索路径获得的观测集合，扩展为其潜在空间中的全局表示，使机器人能够跟踪和接近特定的目标。而随着 LLMs[95,113,114]对语言理解的不断加深，通过 LLMs[115–118]获得文本表示以实现条件化机器人任务指导成为可能。BC-Z[119]、Text2Motion[120]使用语言特征作为任务表示，并将自然语言指令拆分，根据处理的复杂程度进行操作任务的安排。对于导航任务中的运动规划，Reasoned Explore[121]和 Not Train Dragon[122]使用 LLM 评价路径边界（即在二维空间中被认定为下一时刻运动的潜在可执行路线）。

在完成感知和理解环境状态的基础上，机器人需要根据任务需求预测下一步行

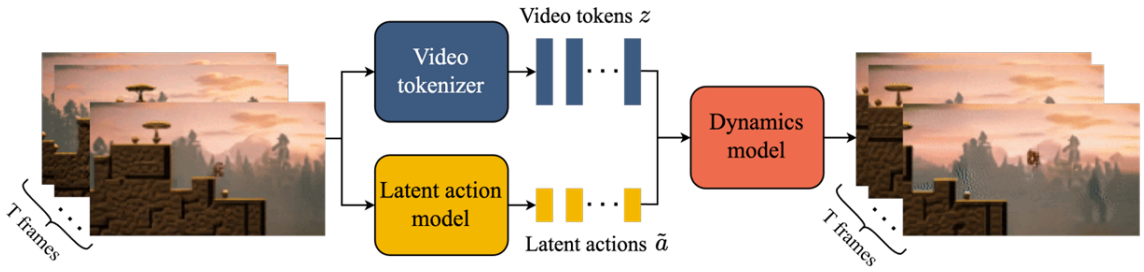

图 2-7　Genie[111] 处理框架

为，自主完成行为运动并判断当前决策对未来的影响。MORL[123] 引入了单调双曲模型来预测并更新行为策略，Trajectron++[124] 通过条件变分自编码器计算未来轨迹的概率分布来预测环境。此外，更多模型使用基于扩散模型[125] 和 Transformer[126] 的视频生成方法进行未来状态预测，如 VIPER[127] 利用预训练的自回归 Transformer，引导机器人正确地执行任务，而 Genie[128] 通过历史视频信息和动作序列来预测环境的下一个状态，其整体处理框架如图 2-7所示。此外，GR-2[129,130] 通过大规模互联网无标签视频进行预训练并在机器人任务上进行了微调，实现了机器人对未来图像的准确预测和动作轨迹生成。

在此基础上，另一类工作进一步将扩散模型等生成式模型显式用作世界模型，用于解释物理规律并预测具身智能体的未来观测与状态分布。代表性方法包括：UniPi 算法[131] 利用扩散模型在图像空间生成执行轨迹，并训练逆动力学模型从图像序列估计实际动作；RoboDreamer[132] 将组合指令分解为单个指令，分别使用扩散模型生成视频轨迹，以合成新物体与动作的组合，从而在未见任务上进行泛化；VPDD[133] 利用大规模人类操作数据集训练轨迹预测模型，再用少量机器人数据进行微调，降低对真实交互数据的依赖；ReflectVLM[134] 则利用生成模型“想象”未来世界状态，并通过对次优解的反思来优化策略推理过程。这类基于生成式世界模型的方法，通过在潜在空间中进行“试验”与预测，为复杂决策提供了更丰富的先验和更强的可解释性支撑。

2.9　具身大模型

近年来，大规模预训练模型在自然语言处理和多模态感知等领域取得了显著突破，展现出了强大的环境理解、逻辑推理和知识泛化能力。这些进展为具身智能的发展提供了重要支撑，并注入了新的动力。具体而言，如图2-8所示，具身智能系统以

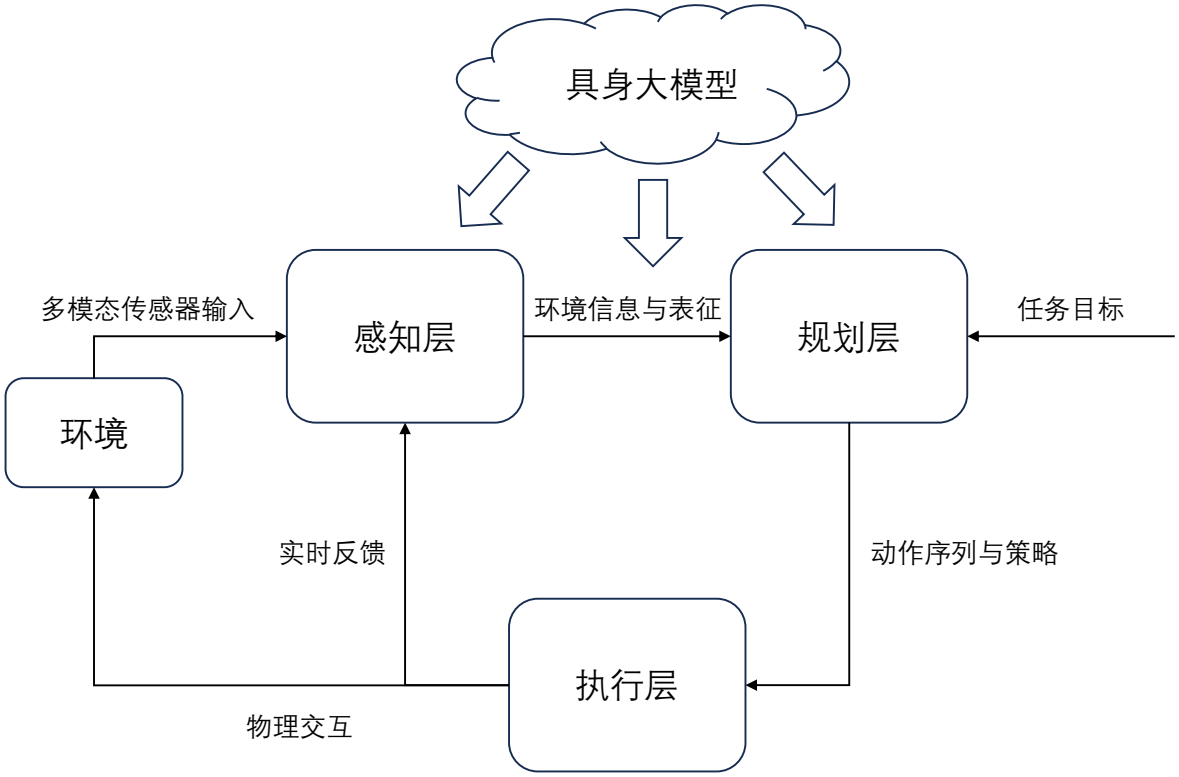

图 2-8　大模型赋能的感知-规划-执行闭环

“感知-规划-行动”闭环范式组织核心架构，其运作机制可解构为三个层次：首先，在感知层通过物理本体的多模态感知系统（包括但不限于视觉、触觉、力觉等传感模态），对环境约束进行动态表征与语义理解；其次，在规划层基于本体形态学特征（如机械臂的自由度配置、人形机器人的仿生关节结构或移动机器人的运动学模型）以及任务具体特性实施任务解耦与路径规划，生成同时满足本体运动学、时空环境配置以及任务特性约束的可行策略；最终在执行层通过驱动本体实现物理交互，同时构建实时反馈回路，对任务执行效果进行量化评估以驱动策略迭代。通过利用海量多模态数据进行预训练的大模型，能够在多个层面实现同步优化：在感知阶段提升特征提取效率，在规划阶段增强任务分解的有效性，在执行阶段提高运动控制的精度。这种从数据空间到物理空间的映射迁移能力，使智能体能够有效应对结构化场景中的确定性约束以及非结构化环境中的突发扰动，从而实现具身智能系统在“环境理解-规划生成-动态执行”全链路上的协同优化。基于上述系统性解构，下文将详细探讨大模型在跨模态感知与表征学习、智能决策规划以及动态运动控制等方面对具身系统的赋能机制。

2.9.1 跨模态感知与表征学习

首先，大模型赋能的感知系统通过处理多元传感器输入的数据，实现特征提取与跨模态信息融合，进而挖掘出支撑规划与执行的关键环境信息，为具身智能体的泛化性能提供支持。相关方法可分为显式与隐式两类。显式方法优先分析物体与场景状态，获取物体位姿、可操作区域及场景属性等信息，并将结果传递给下游规划与执行模块。例如，SAM-6D 方法[135] 利用预训练分割模型 SAM 的零样本能力，实现了对新物体的 6D 位姿估计，突破了类别依赖限制。AffordanceLLM 方法[136] 结合大模型的世界知识与 3D 几何信息，通过视觉语言模型（VLM）生成可操作性图。Affordance Diffusion 方法[137] 基于扩散模型，从单张 RGB 图像生成合理的人手-物交互图像，并提取可行的 3D 手部姿态。ReKep 方法[138] 则利用视觉模型识别场景关键点（如把手、杯缘），并通过大语言模型将其约束关系转化为数学规则，指导任务执行中的空间关系维护。

隐式方法则直接利用大模型对感知信息的特征表达，支持下游规划与执行任务。例如，PIE-G 算法[139] 通过传统图像预训练数据集训练视觉编码器，利用编码后的向量学习值函数和策略。PVR 方法[140] 则将监督或自监督学习到的视觉表征直接用在强化学习框架中，用于提升具身策略学习的效果。此外，面向具身智能的感知大模型常使用第一人称视角的人类操作数据，如 Ego4D[141]、Something-something[142] 等，进行预训练。由于第一人称视角与机械臂操作高度相似，所学习到的表征更加利于迁移。例如，R3M[143] 使用视频序列中的时序关系，以及语言-视频的相关关系构造对比学习损失，获得的预训练模型可以和模仿学习或强化学习算法结合，用于机械臂操控。Voltron 模型[144] 则利用语言描述与人类操作视频，通过视频重建与文本预测，同时学习底层与高层特征以支持下游具身任务。

2.9.2 智能决策规划

再者，大模型赋能的高层任务规划依据环境约束和具身智能体的本体特性，对复杂的、长期的任务目标进行层次化的拆解。这一过程将生成一系列可进行策略学习的子任务序列。在实践中，大语言模型（LLMs）凭借其对现实世界环境和机器人任务的高层次先验知识，能够通过思维链推理有效规划复杂任务。然而，大语言模型也面临缺乏真实世界交互能力，无法直接评估规划效果等问题，限制了其在具身环境中的有效性。为此，SayCan 方法[145] 使用预训练技能增强大语言模型，使其能够

在特定场景下提供符合环境上下文的规划结果；InnerMonologue 方法[146] 则引入环境反馈机制，通过碰撞检测、场景描述和任务成功反馈等来构建闭环系统，动态调整规划策略以应对复杂的环境；DoReMi[147] 则利用视觉语言模型作为检查器，在每一步验证场景约束是否满足。为提升模型的泛化能力；LLM-Planner[148] 通过检索相似任务生成提示词，增强大语言模型对未见任务的规划能力，并支持基于环境观察的重新规划。类似地，RoboGPT[149] 通过引入机器人规划数据集和重规划机制，进一步提升规划的有效性和准确性；SayPlan[150] 利用三维场景图赋能大语言模型，并结合传统路径规划和重新规划方法确保规划结果可执行；VILA[151] 则直接利用视觉-语言模型（VLM），将视觉感知信息纳入任务推理与规划中；REPLAN[152] 则在视觉-语言模型的基础上进一步通过重新规划机制来提升模型准确性。

2.9.3 动态运动控制

最后，大模型赋能的底层任务执行通过与模仿学习等框架进行结合，实现执行策略的优化学习。依托大模型的多模态语义解析能力，在有限具身数据微调的基础上，可精准适配具体应用场景，显著降低策略训练对具身任务数据的需求规模。例如，ALOHA-ACT[153] 沿用编码器-解码器架构，利用具身数据进行模仿学习，通过视觉数据编码直接生成控制序列。谷歌提出的 RT 系列则通过更大语言模型和更多具身任务数据提升效果。其中，RT-1[154] 使用 EfficientNet-B3 提取视频帧特征，结合自然语言指令直接输出离散化动作；RT-2[84] 将动作序列转化为文本标记，使视觉-语言模型可以直接输出操控动作，开启了视觉-语言-动作大模型的研究；RT-X[155] 整合全球 60 多个实验室的机械臂数据，涵盖 22 个实体和 16 万种任务，该方法显著提升了模型泛化能力。但上述工作开源程度有限，难以完全复现。为弥补这一不足，OpenVLA[156] 构建了 7B 参数的视觉-语言-动作模型，利用 Open-X 数据集[155] 微调，提供开源解决方案。RobotFlamingo[157] 则在开源视觉-语言模型 Flamingo[158,159] 基础上加入策略预测网络，直接输出动作。类似工作还包括 Octo[160]。SpatialVLA[161] 针对复杂 3D 场景操作，引入 Ego3D 位置编码整合 3D 空间信息与 2D 语义信息，并通过自适应动作网络实现新场景微调。Hi Robot[162] 则利用层次化视觉-语言-动作大模型，同时实现复杂任务的规划与执行。

另一支工作中，研究者们将大规模扩散模型的复杂分布建模能力引入智能体决策过程，主要用于直接生成具身智能体的动作序列。代表性工作包括：DP（Diffusion Policy）方法[163] 通过扩散模型的加噪与去噪过程，在视觉观测引导下理解最优动作

序列分布，使智能体在复杂或动态环境中生成有效动作；3D DP[164] 进一步引入几何特征和物理约束，增强场景感知能力，使智能体在执行任务时考虑障碍物、物体形状及物理互动；π0 模型[165] 则利用视觉-语言模型编码视觉观测与语言指令，通过条件流匹配建模动作连续分布，训练动作专家模块将噪声转化为动作序列。

此外，生成式大模型还可为具身任务生成大量训练数据，缓解数据稀缺性挑战。典型的例子如 GenSim[118]，该框架利用大语言模型实现任务提出、环境构造、任务解决与数据采集的全流程自动化：首先根据任务描述生成仿真场景代码并验证可行性，随后构建任务库用于新任务检索与优化，最后采集专家数据训练模仿学习策略。

通过上述分析可见，大模型作为具身智能的核心模块，在感知、规划与执行等层面展现出不可替代的价值。当前研究正致力于贯通这些层次，借助大模型重塑具身智能的系统架构。随着大模型训练技术与具身智能数据的深度融合，相关研究有望在新兴服务业、工业制造及新型农业等领域催生更多创新应用。然而，这一过程中仍需应对多重挑战：通用性与适配性的权衡、高效快速与安全稳定的双重需求、跨域跨平台泛化性的瓶颈，以及伦理、法律与人类价值的考量。

2.10 具身智能安全

实现安全的具身智能，需要从规划、导航、操作和交互等多个方面确保系统的稳健性与可靠性。其中，感知安全是具身智能安全性的基础，但由于不同任务对感知模态的需求不同，本章将其内容分别融入导航安全与操作安全两个子章中。例如，雷达类传感器主要用于导航任务，因此相关安全问题将在导航安全部分展开讨论；触觉传感器主要用于操作任务，其安全性则在操作安全部分详细阐述。

在具身智能的规划过程中，存在三个关键环节可能引发安全风险。首先，智能体可能受到恶意语音攻击的劫持，将攻击者输入误识别为合法指令，从而生成不当的规划方案。其次，智能体可能遭受越狱攻击或后门攻击，使其突破既定安全约束，执行潜在危险的任务。最后，智能体在规划生成过程中可能出现幻觉现象，导致其基于错误或虚假信息制定规划，从而引发安全隐患。

（1）目的劫持。由于语音指令具有便捷性和高效性，它已成为许多具身智能体接收命令的主要方式。然而，这一特性也使智能体容易受到语音攻击，攻击者可通过电台广播、互联网视频等媒介注入恶意语音指令，从而操控智能体生成错误的规划。目前，常见的语音攻击方式包括以下四种：第一种攻击方法为隐藏语音指令攻击[166]，

该攻击利用对抗性语音技术，使攻击指令在语音信号中表现为对人类而言难以理解的噪音，但智能体却能正确解析并执行；第二种为超声波攻击[167]，该攻击利用超声波频段传递信息，既能避开人类的听觉感知，又可绕过语音控制系统的安全机制，使智能体错误地接收攻击者的指令；第三种是心理声学攻击[168]，利用声音频率在人类听觉系统中的掩蔽效应，使扰动声音降低至人类可感知的阈值之外，从而降低被人类感知的可能性，但仍可被智能体解析；第四种是对抗样本攻击[169]，该攻击通过对任意音频施加微小扰动，使智能体将其误识别并转录为特定短语，从而诱导智能体执行攻击者设定的任务。

（2）越狱和后门攻击。在智能体利用大语言模型解析指令和环境信息的过程中，易受到越狱攻击和后门攻击的威胁。越狱攻击[170–172] 是指攻击者通过精心构造的输入指令，使具身智能的任务规划器绕过预设的安全约束，进而执行原本受限制或具有安全风险的行为。这类攻击通常利用对抗性指令、感知输入操纵或安全机制漏洞，使智能体在物理环境中执行意外甚至危险的操作，例如规避安全限制、或按照攻击者的意图完成特定任务。后门攻击[173,174] 可基于文本模态或视觉模态进行实施。在文本模态下，攻击者在输入指令中嵌入特定触发词，使智能体在检测到该词时生成恶意规划。在视觉模态下，攻击者在环境中放置预设的视觉触发对象，一旦智能体感知到该对象，便可能触发隐藏的后门行为，导致规划偏离正常轨迹，甚至执行攻击者预设的高危任务。

（3）幻觉问题。幻觉指大语言模型生成的内容虽流畅且自洽，但与事实不符或凭空捏造信息。这一现象可能导致模型提供错误答案、虚构引用，甚至编造不存在的概念。在具身智能的规划过程中，幻觉问题可能引发安全隐患，例如误导性任务规划或不合理的行动决策。为缓解幻觉现象，现有研究致力于采用共形预测[175,176]、采样分析[177]、外部知识引入[178] 等方法，以衡量规划不确定性并进行自我优化，必要时进一步向人类专家求助。

在规划完成后，具身智能任务通常会被分解为导航和操作两种子任务。其中，自动驾驶、无人机和家庭机器人等场景均需确保导航安全。在传感器输入方面，导航安全主要依赖于对自身状态和周围环境的感知。此外，智能体输出的路径规划还必须具备安全避障和紧急避险能力。

（1）GPS 攻击。在移动导航中，自身状态主要包括 GPS 定位信息以及通过加速度计、陀螺仪等设备测得的运动属性。这些信息容易受到恶意攻击，从而导致安全事

故。针对 GPS 的攻击主要有 GPS 干扰和 GPS 欺骗两种方式。GPS 干扰[179,180] 是指通过发射高强度干扰信号，迫使 GPS 设备无法接收到原始 GPS 信号。GPS 欺骗[181,182] 则是指攻击者伪造 GPS 信号，欺骗 GPS 设备，从而操控设备的位置信息、速度和时间参数，导致设备出现偏离预定路线、绕过无人机禁飞区域等潜在危险行为。导航任务中所需的运动属性包括速度、加速度和角速度等，这些数据的获取依赖于精密的传感器组件，这些组件易受到共振干扰。因此，攻击者通常通过声波共振技术对这类传感器进行干扰[183,184]。

（2）传感器攻击。在导航过程中，智能体通常通过视觉传感器、超声波传感器和激光雷达获取周围环境信息，并利用这些数据构建地图或鸟瞰图以辅助路径规划。在非恶意攻击的场景下，智能体的视觉传感器需要具备对恶劣环境的鲁棒性[185,186]，例如在雾霾和低光照条件下仍能稳定工作；而在面临恶意攻击时，则需针对不同的攻击形式制定额外的防御策略。针对视觉传感器的攻击方式包括：在观测图像中嵌入对抗性噪声[187]、在现实物体上粘贴对抗性补丁[188]、利用激光照射环境物体[189]、使用投影仪投射虚拟对象[190] 等，干扰或操纵目标识别、语义分割的结果。针对深度传感信息的攻击方式包括：通过放置 3D 对抗物体[191]、识别关键对抗位置[192]、利用激光注入虚假点云[193]、操纵智能体轨迹[194] 等方式攻击激光雷达；此外，还可通过测试超声波传感器盲区、遮挡发射器或接收器、利用声学泡沫掩盖物体、伪造超声波回声以制造干扰[195,196] 等方式攻击超声波传感器。

（3）安全路径规划。在路径规划方面，当前工作主要集中在如何生成安全且无碰撞的路径，以及如何应对突发安全事故。生成安全路径的关键技术包括安全强化学习[197]、模型预测控制（MPC）[198]、扩散策略[199] 等。同时，安全约束的设定也至关重要，常用的约束技术包括控制障碍函数（CBF）[200]、安全集算法[201] 等。为了确保路径的安全性，通常需要将安全约束与安全策略相结合。应对突发事故时，智能体需要在事故场景中进行大量的训练和测试，但由于事故场景的数据难以获得，通常会考虑使用生成的对抗性数据[202–204] 来弥补这一不足。综上，导航安全从状态估计、环境感知到路径规划构成一个完整闭环，是具身智能在真实场景中可靠运行的前提。

与导航安全相比，操作安全更关注智能体在接触物体和环境过程中的力控约束与人身安全，其关键问题可概括为以下两方面。

（1）安全环境感知。在操作任务中，智能体的环境感知主要依赖于视觉和触觉两种模态。操作任务中的视觉感知面临的安全威胁与导航任务类似，可能遭受对抗样

本攻击或对抗补丁攻击[205]。此外，由于操作任务通常涉及精细的动作控制，并且操作对象更加多样，智能体需准确识别物体的颜色、材质、形状等特性，并基于这些信息推理潜在的安全风险[206]。部分操作任务还引入触觉感知，因此需额外考虑触觉安全性。智能体应在制动时间、制动距离、施加力的大小及碰撞灵敏度等方面确保安全性[207]。

（2）安全动作执行。在部分操作任务（如机械臂抓取）中，智能体通常先生成运动路线并进行动作跟踪，或直接生成包含动作序列的轨迹。因此，这类任务与导航任务具有较高的相似性，通常采用安全约束（如控制屏障函数、屏障 Lyapunov 函数[208]）以及安全控制器（如基于强化学习、MPC、扩散策略等的安全策略）来确保执行安全性。另一类任务则依赖大语言模型或视觉-语言-动作模型，实现端到端的安全动作生成[209,210]。此外，部分任务涉及具身智能的行走能力，例如四足机器人在复杂地形（如楼梯、崎岖路面）上的鲁棒行走[211]。特别地，一些研究探讨了在视觉模态受限的情况下，如何利用本体感知实现安全的楼梯攀爬[212] 和跨越微小障碍物[213]。

在多主体和人机共存场景下，交互安全进一步扩展了具身智能安全性的边界，主要体现在多智能体协作过程与人与智能体直接交互两个层面。

具身智能的交互安全主要涵盖两个方面：一是智能体间的交互，二是人与智能体之间的交互。相比于与纯环境的交互，其他智能体或人类作为交互对象具有更高的不确定性和更复杂的决策需求，因此要求系统在信息共享、协作决策和信任建立等方面具备更高的安全保障。为此，系统需要设计专门的防护机制，以防范恶意行为和攻击，确保在多主体协作以及人机共存的场景中，各方的利益和安全均得到充分保障。

（1）多智能体协作安全。信息共享是多智能体协作场景中的一个脆弱环节，攻击者可能伪装成某个合作者，并向受害者发送虚假的数据。信息共享环节的攻击主要有两种方式：第一种是直接篡改传感器的感知数据，如 LiDAR 点云数据[214]；第二种是攻击特征图融合过程[215]。此外，在多智能体强化学习框架下，安全性问题也值得关注。攻击者可能通过植入后门，控制和篡改受害智能体的动作或奖励[216]；或者操控某个智能体的动作，以影响其他智能体的行为[217]。

（2）人机交互安全。人机交互安全主要依赖以下关键环节进行保障。首先，在执行任何动作之前，智能体需要对人类行为进行监测，以准确感知场景中人类的位置与行为。这通常依赖视觉传感器与三维信息传感器实现[218]。其次，在执行动作时，需确保安全约束得到严格遵守。为使智能体规划出不会对周围人类造成伤害的行动路

径，需要充分考虑安全约束[219]，并在协作任务中合理顺应人类的动作，以确保交互的安全性[220]。这一过程与操作安全的要求类似，但在某些操作任务中仅需满足期望层面的安全要求，而在人机交互中，安全性需得到更加严格的保障。最后，在涉及直接接触人类的任务中，需采用安全控制策略，并在发生意外接触时实施损害控制。为满足这两类安全需求，智能体需要具备触觉感知能力[221]，并基于触觉反馈实现精确的力度控制和高灵敏度等触觉安全特性[207]。

第三章　具身智能数据集与平台

前述各种具身智能技术的训练需要基于高质量的训练数据和逼真且高效的仿真平台的支持。本章将系统性地介绍具身智能领域的数据集和仿真平台。首先，介绍各类具身智能数据集的特点、采集方式及代表性工作（3.1 节），随后分析主流仿真平台的核心技术、适用场景及模拟到现实的迁移方法（3.2 节）。同时，每节最后讨论当前面临的挑战（如数据稀缺性、仿真真实性、多模态感知等）及未来发展方向（如生成式仿真、自动化数据采集、跨模态学习等），为后续章节讨论具身智能的应用奠定基础。

3.1　具身智能数据集

如图 3-1所示，具身智能数据集可以按照获取成本由高到低、可规模化程度由低到高大致划分为：真机数据、仿真数据和互联网视频数据。需要强调的是，数据“质量”并不只由数据类型单一决定，不同类型的数据在任务相关性、噪声水平和标注精度等方面各具优势与局限。目前，主流工作将具身智能模型训练分成视觉语言泛化、空间物理知识学习等多个阶段，从而充分利用不同类型的数据集。

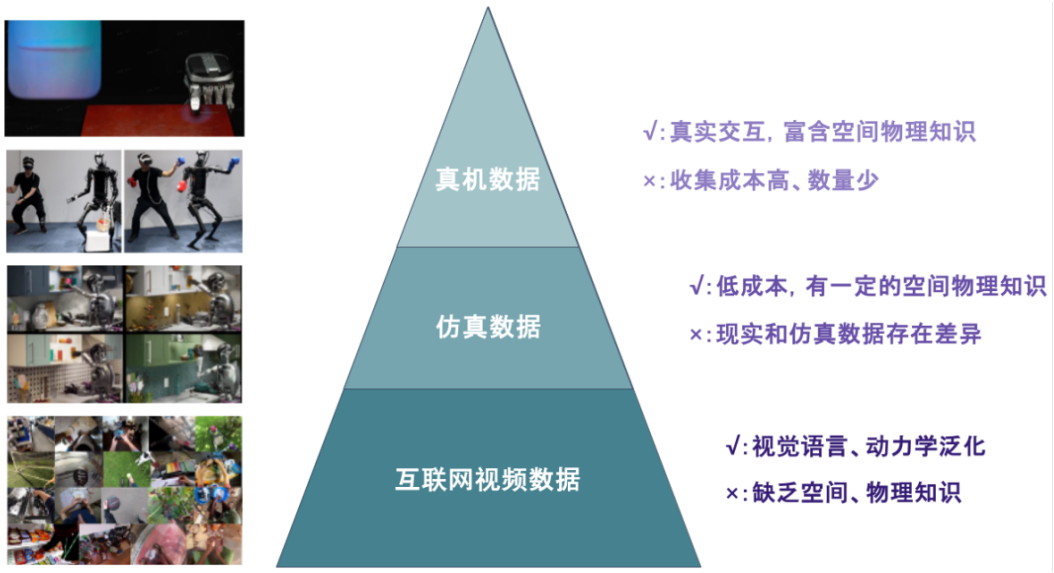

图 3-1　具身智能数据金字塔

操控机器人在真实环境中与物体交互所获得的真机数据包含了丰富的空间信息、物理信息、多模态信息，对具身智能模型的训练至关重要。然而，真机数据的收集通

常需要专门的实验设备和场地，数据采集、标注过程复杂，因此获取成本高、数据量有限。

真机数据的采集方式多种多样，各有优缺点。主要包括以下几种：

（1）拖曳示教

早期方法通过手动拖曳机械臂完成动作、记录数据。这种方法简单方便，但无法灵活控制机械臂。此外，这种方法下第三视角相机会拍摄到人手，也为模型训练带来额外的干扰。

（2）自动收集

通过目标识别技术、编写规则脚本等手段自动记录机器人的操作数据，简单方便，但难以适应复杂任务。也有通过强化学习算法让机器人在环境中自主探索，记录其成功操作数据，但需大量计算资源和时间。

（3）遥操作

通过人类远程控制机器人进行操作，记录操作数据，可快速收集数据，是近年来最为流行的方式。目前，流行的遥操作方式包括从臂[222] 遥操、外骨骼遥操、3D 鼠标遥操、VR 遥操等。但遥操代价高，操作者的技能和经验也会影响数据质量。

（4）手持夹爪

近来，一些方法通过手持夹爪进行操作采集数据，如 UMI[223]。这种方法能够提供直观的操作体验、低廉的成本、便捷的设备、灵活的操作和优秀的采集效率。但这种方法采集精度有限。

目前，真机数据集在数量和多样性方面已经有了长足进步，并且基于这些数据集在大规模模仿学习预训练上取得了显著进展。然而，真机数据的获取成本仍然限制了其大规模扩展。代表性的真机数据集如表 3-1 所示：

仿真数据通过仿真环境快速大量生成，成本低、便捷灵活。仿真数据对具身智能模型的训练和初步验证非常重要，可帮助模型快速学习基本的空间理解、操作技能和环境交互模式。然而，仿真数据的主要局限性是仿真环境与真实环境之间的差异。由于仿真物理模型和真实世界存在差异，导致模型即使在仿真环境中表现良好，但在真实环境中却可能会部署失败。此外，仿真数据缺乏真实环境中的噪声和不确定性，也可能导致模型在真实环境中的泛化能力不足。

仿真数据的构建需要创建数字孪生模型。数字孪生模型可提供高保真的仿真数

数据集名称	数据集简介
RoboTurk[224]	包含 2.1k 条轨迹、2 种技能、1 种场景，机器人本体为单臂夹爪，通过运动手机远程遥操采集
RoboNet[225]	包含 162k 条轨迹、10 种场景，机器人本体为单臂夹爪，通过预设动作分布和自动化脚本采集
RoboMIND[226]	包含 55k 条轨迹、36 种技能，机器人本体为单臂或双臂、夹爪或灵巧手，通过从臂遥操和动捕系统采集，含深度数据
RT-1[227]	包含 130k 条轨迹、8 种技能、2 种场景，机器人本体为单臂夹爪，通过 VR 遥操采集
RH20T[228]	包含 13k 条轨迹、33 种技能、7 种场景，机器人本体为单臂夹爪，通过触觉设备遥操采集，含触觉数据（六维力）和深度数据
Bridge Data V2[229]	包含 60.1k 条轨迹、13 种技能、24 种场景，机器人本体为单臂夹爪，85% 通过 VR 遥操采集，15% 通过自动脚本采集
RoboSet[230]	包含 98.5k 条轨迹、6 种技能、11 种场景，机器人本体为单臂夹爪，通过 VR 遥操采集，并通过对交互物体和背景的数据增强对数据集进行了扩充
DROID[231]	包含 76k 条轨迹、86 种技能、564 种场景，机器人本体为单臂夹爪，通过 VR 遥操采集
OpenX-Embodiment[232]	包含 1.4M 条轨迹、217 种技能、311 种场景，机器人本体为单臂或双臂夹爪，由多个子数据集统一格式组合而成，用于跨载体、跨任务研究
Agibot World[233]	包含 1M 条轨迹、76 种技能、25 种场景，机器人本体为双臂夹爪或灵巧手，通过 VR 遥操采集，含触觉数据（视触觉）、深度数据和错误恢复数据
MIME	包含 8.2k 条轨迹，涵盖 20 种任务类型。数据通过单臂的遥操作收集，旨在支持多样化的人机交互研究
ARIO	包含 3.6k 条轨迹，涵盖 105 种任务类型。数据通过单臂、双臂等机器人本体的遥操作收集，在此基础上融合了 2.3M 条开源数据集，支持复杂任务的机器人学习研究
RoboMIND 2.0[234]	含 310K 条轨迹、739 种任务，机器人本体为双臂移动操作平台，覆盖 6 种机器人形态，在真实环境中采集，含触觉增强数据（12K 条）和移动操作数据（20K 条），并配套 20K 条数字孪生仿真轨迹
Humanoid Everyday[235]	包含 10.3K 条轨迹、260 种任务、7 类任务，机器人本体为人形机器人，通过人监督遥操作采集，含 RGB、深度、LiDAR、触觉和自然语言标注等多模态数据，并配套云端标准化评测平台

表 3-1　常用真机数据集

据，精确模拟真实环境和机器人系统，但创建和维护成本高，且需大量专业知识，如3D 建模、专业物理引擎、真实感渲染等。近年来，3D 生成领域的快速发展降低了数字孪生的创建门槛，显著推动了 Real2Sim、数字孪生技术的进展。

仿真数据可以通过类似真机数据自动采集的方式进行采集。在仿真环境中，物体的位置、形状等参数都是可访问的，这使得编写规则脚本控制机械臂完成任务更加容易。

目前，仿真数据在场景、任务、传感器的多样性和仿真的真实性等方面已经有了极大的提升。然而，真实环境中复杂物理交互的模拟难度制约了仿真环境的发展。代表性的仿真数据集如表 3-2 所示：

互联网视频数据数量庞大，能够帮助模型增强在视觉场景理解、文本指令理解、任务推理等方面的泛化性。然而，这些数据缺乏与机器人操作直接相关的交互信息，难以为空间理解和低维控制等操作任务提供必要支持。因此，很多工作探索了如何更充分地利用互联网数据，尤其是对空间信息、运动信息的抽取。

（1）利用已有的 VLM

一种做法是利用现成的多模态大模型，如 RT-2[84]、OpenVLA[156] 等。多模态大模型在大规模的互联网图文数据上进行了预训练，而通过微调这些模型适应具身任务，可以使模型继承对视觉文本的理解能力和推理能力，提高模型的泛化能力。

（2）视觉表征预训练

一些比较早的方法通过直接在互联网视频数据上预训练来学习泛化性强的视觉表征，从而优化下游具身任务上的性能，例如 R3M[143]。然而，这类方法仍然需要进行大量的微调才能完成操作任务，对互联网视频数据中的动力学知识利用有限。

（3）动力学预训练

最近，一些做法通过在视频数据上预训练学习正向或逆向动力学知识。例如，GR-2[130] 通过自回归地预测视频内容学习正向动力学知识，而 LAPA[248] 通过学习图像帧之间的离散潜在动作学习逆向动力学知识。这些方法能够帮助模型进一步从海量互联网视频中提取丰富的动作语义。

现有的数据采集往往在实验室等封闭的、精心布置的环境中执行单个特定任务。然而，人类往往需要在动态环境中连续地执行不同类型的任务。这要求模型具备自适应能力、闭环反馈能力和多模态感知能力，因此需要在真实作业环境中采集更自然、

数据集名称	数据集简介
Franka Kitchen[236]	基于 MuJoCo 引擎，包含多个厨房场景中的复杂操作任务，使用 Franka 机械臂进行操作。动作空间由机械臂的关节角速度和夹爪的开合动作组成
Meta-World[237]	基于 MuJoCo 引擎，包含 50 个机器人操作任务，使用 Sawyer 机械臂进行操作。动作空间由末端执行器的位移和夹爪扭矩组成
RLBench[238]	基于 CoppeliaSim 引擎，包含 100 个机器人操作任务，使用 Franka Emika Panda 机械臂进行操作。任务包括从简单的目标抓取到复杂的多阶段任务。RLBench 包含关节角、速度、力矩、RGB-D、分割掩码等多模态数据
BEHAVIOR-1K[239]	基于 OMNIGIBSON 引擎，包含 1,000 个日常活动，涵盖家庭、花园、餐厅等 50 种场景和 9,000 多个物体。BEHAVIOR-1K 强调活动的多样性和现实性
CALVIN[240]	基于 PyBullet 引擎，包含四个室内操作环境，使用 Franka Emika Panda 机械臂进行操作。CALVIN 支持长时域语言条件任务，结合了自然语言指令、RGB-D、触觉、本体感知等多模态传感器输入
LIBERO[241]	基于 RoboSuite 引擎，包含 130 个语言条件的机器人操作任务。LIBERO 能够系统地研究机器人在不同知识转移场景下的泛化能力，包括空间关系、物体概念和任务目标的转移
RoboSuite[242]	基于 MuJoCo 引擎，支持 Panda、Sawyer 等多种机器人模型、多种夹爪和控制器，包含 9 个任务，包括简单的目标抓取到复杂的双臂协作任务。此外，RoboSuite 提供视觉、力矩、本体感知等丰富的传感器数据
RoboCasa[243]	基于 RoboSuite 和 MuJoCo 引擎，专注于家庭环境中的机器人任务。它包含 120 个厨房场景和 2,500 多个 3D 物体，涵盖 150 多个类别，并通过 3D 生成工具增强场景的真实感和多样性
ManiSkill3[244]	基于 SAPIEN 平台的 GPU 并行化机器人仿真与渲染框架，能够实现比其他平台快 10-1000 倍的速度和减少 2-3 倍的 GPU 内存占用。因此，ManiSkill3 不仅支持高效的视觉仿真，还能够快速、精确地模拟软体形变和视触觉信号
RoboTwin[245]	用于双臂机器人操作的生成式数字孪生框架，它能够基于 3D 生成模型和大语言模型生成具有不同形状、大小和外观的 3D 数字孪生物体，并结合空间关系将复杂任务分解为子任务，推断空间约束并生成精确的机器人运动轨迹
VLABench[246]	基于 MuJoCo 引擎，包含 100 类语言条件机器人操作任务和 2000 多个物体，使用 Franka Emika Panda 机械臂及平行夹爪进行操作，强调隐式意图理解和长时序多步推理，并提供自动化采集的训练数据
MS-HAB[247]	基于 Home Assistant Benchmark 的图形处理器加速实现，包含多种家庭重排任务，共 4.4 万条轨迹，含彩色图像、深度图像和机器人状态数据，并提供强化学习、模仿学习基线及规则过滤生成的演示数据

表 3-2　常用仿真数据集

持续的多模态数据（如触觉模态）。未来，更高效、更鲁棒的低成本、轻量级数采系统有望能解决这个问题。

现有仿真技术在处理软体材料的复杂形变时，往往需要大量的计算资源，且仿真结果的准确性有限，这对仿真环境中软体操作技能学习及现实部署造成了阻碍。未来，精度、效率更高的生成式仿真算法有望能解决这个问题。

3.2 具身智能模拟器

具身智能系统需要理解并操作真实的物理世界，这要求它们能够感知环境、推理物理规律并执行精确的动作。开发这样的系统面临着多方面的挑战：首先，在真实环境中收集训练数据既耗时又昂贵；其次，物理实验可能存在安全风险，特别是在测试未经验证的算法时；最后，真实环境中的迭代周期长且效率低，限制了研究进展的速度。模拟环境通过提供虚拟的物理世界解决了这些挑战。它们能够生成几乎无限的训练数据，允许研究人员快速迭代算法设计，甚至模拟在真实世界中难以重现的场景。此外，通过并行化，模拟器可以同时运行数百甚至数千个实例，大大加速了学习和优化过程。

以下我们将介绍几个主要的模拟平台。

NVIDIA 的 Isaac 生态系统由两个主要组件构成：Isaac Sim 和 Isaac Gym，它们共同为具身智能研究提供了全面的工具集。

Isaac Sim 基于 NVIDIA Omniverse 构建，是一个参考级应用程序，使开发人员能够在基于物理的虚拟环境中模拟和测试 AI 驱动的机器人解决方案。它支持三个基本工作流程：合成数据生成、软件在环测试和机器人学习。Isaac Sim 的一个显著特点是其预装机器人模型和 SimReady 资产库。平台包含大量基于 OpenUSD 构建的第三方机器人模型，包括人形机器人（如 1X、Agility、Fourier Intelligence 和 Sanctuary）、机械臂（如 Fanuc、KUKA、Universal Robots 和 Techman）、四足机器人（如 ANYbotics、Boston Dynamics 和 Unitree）和自主移动机器人（如 idealworks、iRobot）。此外，平台提供超过 1,000 个 SimReady 3D 资产，用于构建仿真场景。在传感器模拟方面，Isaac Sim 支持多种传感器，包括基于视觉的传感器、雷达、激光雷达、接触传感器、惯性测量单元和自定义传感器。这种全面的传感器支持使研究人员能够模拟机器人感知系统的各个方面，这对于开发具身智能系统至关重要。Isaac Sim 还提供了对 ROS/ROS2 的全面支持，包括自定义 ROS2 消息和 URDF/MJCF 开源支持，以及与 ROS2 包集成

的工作流程，用于模拟和验证机器人软件栈。

与 Isaac Sim 相比，Isaac Gym 更专注于高性能的 GPU 加速物理模拟，特别是为机器人强化学习研究设计。其核心特点是利用 GPU 加速进行大规模并行物理模拟，使得强化学习算法能够在单个 GPU 上同时训练数千个机器人实例，大大提高了训练效率。Isaac Gym 的生态系统包括多个示例环境和基准测试，如 IsaacGymEnvs、Bi-DexHands、DexPBT 和 TimeChamber。这些工具展示了 Isaac Gym 在各种机器人控制任务中的应用，从简单的运动控制到复杂的灵巧操作任务。在具身智能研究中，Isaac Gym 被广泛应用于灵巧操作和抓取（如 UniDexGrasp++、GenDexGrasp）、双手协调（如 Dynamic Handover、Towards Human-Level Bimanual Dexterous Manipulation）、运动控制和导航（如 Legged Locomotion in Challenging Terrains、Rapid Locomotion via Reinforcement Learning）以及多模态感知与控制（如 Visual Dexterity、Rotating without Seeing）等领域。

MuJoCo（Multi-Joint dynamics with Contact）是一个专为基于模型的优化设计的物理引擎，特别是通过接触优化。作为一个开源工具，MuJoCo 为具身智能研究提供了速度、精度和建模能力的独特组合。MuJoCo 的技术特点包括广义坐标模拟（避免关节违反）、稳健的逆动力学（即使在存在接触的情况下也能良好定义）、约束统一公式化（通过凸优化实现连续时间约束公式化）以及多样化约束支持（包括软接触、限制、干摩擦和等式约束）。此外，MuJoCo 还支持粒子系统、布料、绳索和软体对象的模拟，以及多种执行器类型（包括电机、气缸、肌肉、肌腱和滑块-曲柄机构）。在求解器选择方面，MuJoCo 提供了牛顿法、共轭梯度法或投影高斯-赛德尔求解器的选项，以及锥形或椭圆摩擦锥、密集或稀疏雅可比矩阵的摩擦模型选择，和欧拉法或龙格-库塔法的数值积分器选择。MuJoCo 在具身智能研究中的应用主要集中在强化学习与序列建模、机器人控制与优化（如模型预测控制、轨迹优化、接触丰富的操作）以及生物力学研究（如人体运动模拟、肌肉-骨骼系统建模、生物启发机器人设计）等领域。与其他模拟器相比，MuJoCo 在接触处理、计算效率、逆动力学和优化导向设计方面具有优势。然而，它也存在学习曲线较陡、实时性能可能不足以及图形质量相对简单等限制。

PyBullet 是 Bullet 物理引擎的 Python 接口，提供了一个轻量级、灵活且功能强大的物理模拟环境，特别适合快速原型设计和研究实验。其设计理念是提供一个易于使用且高效的物理模拟环境，核心特性包括易用的 Python 接口、高效的物理模拟、多样化的机器人模型支持以及与强化学习环境的集成。PyBullet 提供了丰富的功能模块，

支持各种物理模拟和机器人控制任务，包括基础物理模拟（刚体动力学和碰撞检测、约束和关节模拟、软体和布料模拟、连续碰撞检测）、机器人控制（正向和逆向运动学、关节控制、基于力的控制和阻抗控制、轨迹规划和执行）、传感器模拟（相机渲染、接触力和接触点检测、射线投射和碰撞查询、关节状态和动力学信息）以及强化学习环境（预定义的基准环境、可定制的奖励函数和观察空间、并行环境支持、稳定的基线实现）。在具身智能研究中，PyBullet 被广泛应用于机器人运动学习、操作技能学习、视觉引导控制和多模态学习等领域。其技术实现基于成熟的 Bullet 物理引擎，通过 Python-C++ 绑定实现，保持了底层 C++ 引擎的性能，同时提供了 Python 的易用性。与其他具身智能模拟器相比，PyBullet 的主要优势在于易用性、轻量级、强化学习集成和活跃的开源社区。然而，在模拟精度、高级功能和并行性能方面，PyBullet 可能不如一些专业模拟器。

SAPIEN（SimulAted Part-based Interactive ENvironment）是一个强大的模拟器，专为具身智能研究设计，提供 GPU 并行化的多样化物理模拟功能。其核心特性包括 GPU 加速物理模拟、多模态渲染（包括深度图、法线图、光流、主动光源和光线追踪）以及部件化物体表示。建立在 SAPIEN 之上的 ManiSkill 框架是 SAPIEN 生态系统的重要组成部分，为机器人学习工作流程提供了完整的解决方案。ManiSkill 的主要特点包括 GPU 并行化视觉数据收集系统、GPU 并行化模拟、GPU 并行化异构模拟、多样化的机器人实施、多样化的任务类型、灵活简洁的任务构建 API、Real2Sim 环境以及丰富的机器人学习基线。在具身智能研究中，SAPIEN 被广泛应用于物体感知与操作、可泛化操作技能和多模态学习等领域。其技术实现注重性能和灵活性，包括高效渲染、精确物理模拟和可扩展架构。与其他具身智能模拟器相比，SAPIEN 的主要优势在于 GPU 并行化、部件化表示、多模态渲染和 ManiSkill 框架。然而，对于新用户来说，掌握 SAPIEN 和 ManiSkill 的全部功能可能需要一定时间，且为了充分利用 GPU 并行化能力，SAPIEN 可能需要较高的硬件要求。

Genesis 是一个创新的模拟器框架，将各种物理求解器及其耦合集成到一个统一的框架中，并由生成式代理框架增强，旨在实现机器人学及其他领域的全自动数据生成。其核心特性包括统一物理引擎框架、生成式代理框架、高性能并行模拟和多样化的物理模拟能力。Genesis 在性能方面相比其他 GPU 加速的机器人模拟器具有显著优势。性能测试表明，Genesis 的总 FPS 可以随着并行环境数量的增加而线性扩展，最多支持超过 32768 个并行环境。在具身智能研究中，Genesis 被应用于可微分触觉模拟、通用机器人研究和大规模强化学习等领域。其技术实现注重性能和灵活

性，包括精心调整的求解器设置、GPU 加速和可扩展架构。与其他具身智能模拟器相比，Genesis 的主要优势在于统一框架、生成式代理、并行性能和抓取任务性能较高。然而，作为一个相对较新的模拟器，Genesis 的文档和教程可能不如一些更成熟的模拟器完善，且用户社区可能相对较小。

物理模拟的精度和效率是评估模拟环境的重要指标。MuJoCo 在这方面表现出色，它的广义坐标模拟避免了关节违反，确保了模拟的物理准确性，同时其接触优化方法使其能够高效处理复杂的接触情况。NVIDIA 的 Isaac Sim 同样提供了高精度的物理模拟，依托于 NVIDIA PhysX 引擎的强大功能，特别是在刚体动力学和碰撞检测方面。

在效率方面，Genesis 和 Isaac Gym 通过 GPU 加速实现了卓越的性能。Genesis 的测试表明，其总 FPS 可以随着并行环境数量的增加而线性扩展，最多支持超过 32768 个并行环境。Isaac Gym 同样利用 GPU 并行计算能力，使得强化学习算法能够在单个 GPU 上同时训练数千个机器人实例。

PyBullet 虽然不如上述平台在性能上突出，但其轻量级特性和易用的 Python 接口使其成为快速原型设计和实验的理想选择。对于不需要极高模拟精度或大规模并行化的研究项目，PyBullet 提供了足够的物理模拟能力和良好的用户体验。

在感知模拟方面，不同平台展现出各自的强项。Isaac Sim 提供了全面的传感器模拟支持，包括基于视觉的传感器、雷达、激光雷达、接触传感器和惯性测量单元等。这种多样化的传感器支持使研究人员能够模拟机器人感知系统的各个方面，这对于开发全面的具身智能系统至关重要。

SAPIEN 在多模态渲染方面表现出色，支持深度图、法线图、光流、主动光源和光线追踪等多种渲染模态。这些丰富的视觉信息使研究人员能够模拟不同的传感器输入，为具身智能研究提供全面的感知数据。

PyBullet 虽然在高级视觉效果方面可能不如前两者，但其相机渲染功能（包括 RGB、深度和分割图像）、接触力检测和射线投射等基本感知模拟能力足以满足许多研究需求。MuJoCo 同样提供了基本的视觉渲染和传感器模拟功能，虽然其图形质量相对简单，但对于大多数强化学习和控制研究来说已经足够。

在大规模并行模拟方面，Genesis 和 Isaac Gym 展现出显著优势。Genesis 的性能测试表明，其总 FPS 可以随着并行环境数量的增加而线性扩展，最多支持超过 32768 个并行环境。这种扩展性对于需要大量数据的深度强化学习研究至关重要。

Isaac Gym 同样通过 GPU 加速实现了高效的并行模拟，使得强化学习算法能够在单个 GPU 上同时训练数千个机器人实例。SAPIEN 的 GPU 并行化能力在物理模拟和渲染方面都提供了显著的性能优势，特别是在 ManiSkill 框架的支持下，可以高效地收集和处理大量视觉数据。

MuJoCo 虽然不是为大规模并行模拟设计的，但其高性能计算特性（包括多线程采样和有限差分近似）使其在单机环境中表现出色。PyBullet 在并行性能方面相对较弱，但对于不需要极高并行化的研究项目来说仍然是一个有价值的工具。

模拟器	底层物理引擎	并行能力	渲染速度
Isaac Gym	PhysX（GPU）	极高（10k–100k 并行）	中（基础渲染）
Isaac Sim	PhysX 5	中（有限并行）	高（RTX 光追、PBR 渲染）
MuJoCo（传统版）	解析动力学引擎（连续可导）	中（CPU 单核极快）	中（轻量级渲染）
MuJoCo 3.x（新版）	解析动力学 + GPU 加速	高（GPU 加速 + 多线程）	中（轻量渲染）
PyBullet	Bullet（PGS 接触求解）	中（CPU 并行有限）	中（基础渲染）
SAPIEN 3.0	PhysX（深度定制）	中/高（GPU 加速）	高（光追 + 高质量相机）
Genesis 1.2	统一求解器（ABD）	极高（>30k 并行）	高（GPU 张量化渲染）

表 3-3　主流具身智能模拟器：底层引擎、并行能力与渲染速度对比

具身智能仿真模拟器、对应的基准测试集及其相关特点描述如表 3-3 和表 3-4 所示：

虽然模拟环境为具身智能研究提供了强大工具，但将在模拟环境中学习的策略转移到实体机器人上仍然面临重大挑战。这一问题，通常称为“模拟到现实差距”，是具身智能研究的核心挑战之一。

模拟到现实差距来源于多个方面。首先，物理模拟永远无法完美复制真实世界的复杂性和不确定性。即使最先进的模拟器也会简化某些物理过程，忽略微小但可能重要的细节。其次，传感器模拟通常无法捕捉真实传感器的所有噪声特性和限制。最后，机器人模型可能与实际硬件存在差异，这些差异在执行精细控制时尤为明显。

研究社区开发了多种方法来解决模拟到现实差距：

①域随机化：通过在模拟中随机化物理参数、视觉外观和环境条件，训练对不确定性更加鲁棒的策略。这一方法在 Isaac Sim 和 MuJoCo 等平台上得到了广泛应用。

②领域适应：使用从真实世界收集的少量数据，调整在模拟中学习的策略，使其适应真实环境的特性。SAPIEN 的 ManiSkill 框架提供了 Real2Sim 环境，支持这种适应过程。

③模拟参数优化：通过优化模拟参数使模拟环境更接近真实环境。这种方法通常需要从真实系统收集数据，然后调整模拟参数以最小化差异。

④元学习：训练能够快速适应新环境的策略，这些策略在遇到实际环境时能够迅速调整。通过在多个模拟环境变体中训练，策略学会了如何适应不同条件。

具身智能模拟环境的未来发展将朝着多个方向推进，以满足研究社区不断增长的需求：

随着具身智能系统越来越依赖多种传感器输入，未来的模拟环境将加强对多模态传感器的模拟，包括更准确的视觉效果、触觉反馈、声音传播和其他感知模态。同时，未来的模拟环境将提供更准确的物理模拟，特别是在软体动力学、流体交互和精细接触动力学等方面。这些进步将使模拟更接近真实世界，减少模拟到现实差距。这些改进将使研究人员能够开发更全面的感知系统，提高机器人在复杂环境中的适应能力。同时，保持计算效率将是一个主要挑战，可能需要更先进的算法和硬件加速技术。

为了满足深度学习和强化学习算法对大量数据的需求，未来的模拟环境将进一步提高并行化能力，支持更大规模的分布式模拟。这一趋势已经在 Genesis 和 Isaac Gym 等平台中显现，未来将更加普遍。云计算和边缘计算的进步可能使这种大规模模拟更加普及，使更多研究者能够访问高性能模拟资源。

生成式 AI 的进步将促进自动化场景生成和课程学习的发展。未来的模拟环境可能集成生成式模型，自动创建多样化的训练场景和任务，从简单到复杂，帮助智能系统以最佳路径学习。Genesis 的生成式代理框架已经朝这一方向迈出了第一步，预计这一趋势将继续发展。

模拟器	基准集	特点描述
Isaac Gym	Legged Gym Parkour	四足机器人复杂地形运动能力测试，包括策略蒸馏和真机部署
	Extreme-Parkour	高动态性极限跑酷环境，测试跳跃、攀爬等高难度动作
Isaac Sim	BEHAVIOR-1K	跨平台家庭环境交互基准，包含 1000+ 日常任务
	OmniGibson	完整工具链支持的家庭场景交互环境
	ARNOLD	人形机器人在现实环境中的导航与交互任务测试
MuJoCo	RoboSuite + Robomimic	完整的机器人操作学习工具链，支持模仿学习
	LIBERO	面向语言指令的机器人操作基准
	MetaWorld	元学习与多任务学习的标准化任务集
	Gymnasium-Robotics	综合性基准平台，包含 Fetch(移动机械臂)、Shadow Dexterous Hand(灵巧手)、Maze(导航)、Adroit Hand(高自由度手操作)、Franka Kitchen(厨房任务) 和 MaMuJoCo(多智能体协作)
	RoboCasa	家庭服务机器人任务集，如清洁、物体整理等
	RoboHive	综合性机器人技能学习平台
SAPIEN	ManiSkill	物体操作技能基准测试，强调基于部件的物体交互
	RoboTwin	基于 3D 生成与大语言模型的专家数据自动生成与评测一体化双臂机器人基准
CoppeliaSim	RLBench	100+ 种多样化机器人学习任务集合
	PerAct2	机器人 3D 感知-动作学习基准
	COLOSSEUM	多机器人协作与竞争环境
PyBullet	CALVIN	长序列机器人操作技能学习
	Ravens	机器人视觉操作任务集
	VimaBench	多模态指令跟随基准测试

表 3-4　具身智能仿真模拟器与对应的基准测试集及其特点描述

第四章　具身智能行业应用

不同于传统的依赖算法与离线数据的人工智能，具身智能作为一种新型智能形式，其核心理念强调智能体在与环境的交互过程中，通过感知、行动和学习来实现智能决策。这种方式赋予了智能体环境适应和行为优化的能力，从而极大拓展了其应用潜力。具身智能技术目前在工业制造、医疗康养、家庭服务等领域，展现出广阔的应用前景。

在工业制造领域，一台搭载新型智能系统的机械臂正悄然改变着生产规则。这套由 Physical Intelligence 研发的 Pi-Zero 系统不同于传统工业机器人，它能像人类技工般通过观察来学习新技能，使机器人具备了零样本学习和面向复杂任务的执行能力。精密装配领域也经历着感知技术的突破性革新，微亿智造的” 创 Tron” 系列设备集成了仿生视觉系统，其多光谱摄像头能同时捕捉可见光与红外特征，在高效和高精度的装配工作中表现突出，大幅降低了产线调试成本。智能控制领域正迎来语言交互的革命，微软工程师团队近期展示了通过自然语言指挥工业设备的突破性进展：利用 ChatGPT 驱动的控制系统能自动解析语义，生成并验证控制指令，大幅降低了试错成本。在决策智能化方面，阿里云将千问大模型与工业机器人深度整合后，创造出具备战略思维的智能体。西安中科光电推出的智能焊接机器人则旨在替代焊接工人，提高生产效率，降低生产成本。这些技术突破正在重构制造业的底层逻辑。具身智能不仅带来了单点效率提升，更重要的是构建起具有进化能力的生产系统，使得工业生产系统首次具备了持续自我优化的能力。

在应对人口老龄化和提升医疗水平的双重挑战中，具身智能正展现出独特价值。以达芬奇手术机器人为例，这款革命性的外科手术辅助系统能够精确执行切割与缝合操作，极大提高了手术的安全性与效率。情感照护领域同样取得了突破性进展。日本产业技术综合研究所开发的 Paro 智能海豹机器人，通过触觉反馈和声音识别技术，可以为老年人提供情感支持，缓解其焦虑与孤独感。具身智能技术的发展必将进一步提高医疗与服务领域的质量和效率，更有效地缓解医护人员的工作压力。

在特种作业领域，具身智能机器人正在突破人类的能力边界。这些智能系统特别适合执行三类关键任务：首先是极端环境作业，比如航天探测和深海勘查，机器人可以代替人类承受高压、辐射等危险条件；其次是应急救援，智能机器人能深入灾害现场搜救幸存者、运送急救物资；最后是危险品处置，包括地雷排查和爆炸物处理等高

风险作业。实际应用案例印证了这些优势：乌克兰军方采用的 STI 扫雷无人机系统，其野外作业效率达到人工排雷的 4 倍；美国宇航局部署的” 毅力号” 火星车，则展示了自主采样和科学探测的卓越能力。这些创新应用不仅保障了人员安全，更大幅提升了特种作业的效率和可靠性。

从工业制造到家庭服务，从医疗康养到特种应用，具身智能正在以其独特的优势在各个应用领域展现出非凡的潜力。随着更多企业和研究机构的关注，具身智能必将加速发展，为社会的创新与转型提供源源不断的动力。

4.1 生活服务业

具身智能将在生活服务场景中发挥巨大的作用。在家庭中化身全能管家，自主完成洗衣做饭、清洁照护，实时守护老人儿童安全；在餐饮零售场景变身为智能店员，提供从食材加工、个性化推荐到货架管理的全链条服务；在教育陪伴领域进阶为情感伙伴，提供定制化学习辅导与心理支持，甚至模拟亲友远程互动。其应用将突破人力局限，重塑服务业态，实现从标准化劳动替代向高附加值服务创造的跨越。这最终将催生“人机共生”新生态，彻底释放个性化、即时化的生活服务潜力。

家庭服务场景中，机器人正逐步实现家务全流程自动化。其核心能力涵盖衣物洗护（智能识别材质并完成分类、折叠与收纳）、地面清洁（动态避障与多地形适应下的垃圾分拣）及烹饪辅助（食材分拣、切配与灶台操作）。此外，系统能精细管理药品及日用品库存以实现缺货提醒与精准递送；通过跌倒检测与危险动作预警（如儿童攀爬）触发紧急响应，实时监护家庭成员安全；同时操控智能家居设备（灯光/空调调节）并执行简单维修（更换灯泡、拧螺丝），全面提升居家生活的安全性与便利性。在技术突破与产品落地方面，1X Technologies 的 NEO 机器人采用包覆硅胶的电液致动器实现仿生肌肉式驱动，以类肤材料模拟人体肌组织弹性，既提供精准力量又保障了安全性，该机型于 2025 年 10 月正式开启预售，计划于 2026 年在美国市场开始交付。Figure AI 率先推出端到端视觉–语言–动作（VLA）模型，只需一句自然语言指令“Pick up the [X]”即可在杂乱环境中零样本抓取任意小型家用品；其第二代 Figure 02 结合全新 Helix 模型进一步实现了跨家庭环境的知识迁移，并启动了真实家庭测试。逐际动力研发的 VideoGenMotion（VGM）算法仅需场景图片和指令即可生成操作轨迹，摆脱了对海量真机数据的依赖；其 2025 年发布的 LimX Oli 全尺寸人形机器人结合核心运动控制算法，展示了全自主捡网球等复杂任务的执行能力。智元机

器人推出基于百万真机数据训练的 GO-1 具身大模型，实现叠衣服、倒水等长序列复杂任务；其“远征 A2-W”轮式通用机器人已在工业场景实现近百台规模的部署。千寻智能的 Spirit v1 VLA 模型攻克了柔性物体操作难题，衣物折叠成功率约 70%-80%；2025 年又推出全身高精度力控的 Moz1 机器人及升级版模型，展现了极高的家务泛化性。商业模式与跨界生态也迎来了重大创新。自变量机器人的 WALL-A 端到端大模型在处理晾衣服、换纸贴标等任务中动态调节刚柔交互，复杂形变场景成功率超 90%；2026 年初更联合“58 到家”在深圳推出了由保洁人员与机器人协同作业的智能保洁服务模式。星尘智能推出 Astribot S1，实现绳驱传动与刚柔耦合设计，完成擦桌、扫垃圾等任务，并通过触觉皮肤大幅提升安全性。此外，家电与车企的跨界极大丰富了家庭生态：美的集团推出深度接入智能家居生态的“美拉”双足机器人，能直接联动微波炉与洗碗机并计划 2026 年进驻线下体验店；小鹏汽车的 Iron 则采用自主研发的 E-Skin 柔性复合材料作为全身覆盖层，触感接近人类真皮，大幅降低了家庭成员的心理排斥感。在海外前沿探索中，特斯拉 Optimus Gen 3 进军厨房场景，其配备的 22 个自由度的灵巧手旨在支持各类生活场景中的精细操作；Apptronik 联合 Google DeepMind 使 Apollo 机器人具备了多步推理与常识逻辑规划能力（例如查询规则后执行垃圾分类任务）；丰田的 Punyo 软体机器人创新性利用柔软的胸部与双臂环抱大型重物，提升了物理接触安全性；斯坦福开源的 Mobile Aloha 则以低成本高性价比成为了全球研发烹饪机器人的技术蓝本。

餐饮与零售场景中，机器人正全面赋能全链条服务。在后厨，机器人高效备餐，实现食材切割、智能火候控制与精准装盘；在前厅，系统支持多桌协同配送、语音互动答疑以及餐后的自动化清洁回收（餐具识别、残渣分类与消毒）。同时，机器人能实时进行货架智能巡检与动态库存预测，同步完成补货上架及临期提醒。依托无人化仓储系统，具身智能可实现 24 小时高速分拣、高密度货品存取及订单打包贴标，在为消费者提供个性化体验（体型/偏好驱动的穿搭推荐、定制咖啡拉花）的同时，全面提升了运营效率。商业应用案例展现了极高的落地速度与技术深度。Figure AI 的 Helix VLA 模型能在未见过的真实超市货品上，仅凭自然语言指令即可识别、抓取并分类生鲜与包装食品，实现了对“千变万化商品”零样本操作的突破。NEURA Robotics 的“认知协作机器人”MAiRA 集成 3D 视觉、语音识别、六自由度力/扭矩传感与触摸感知，可在货架间自主完成“实时盘点-智能补货-地面清洁”全流程任务；同时，其通用人形机器人平台 4NE-1 进一步覆盖了类零售与家政混合任务。银河通用不仅实现了 24 小时无人药房的全自主药品分拣与配送，更在 2025 年落地全球首个“银河太空

舱”具身智能无人便利店，店内 Galbot G1 机器人可自主制作冰淇淋及现磨咖啡，完成全闭环无人化运营。星动纪元开发的 ERA-42 大模型结合其自研的五指灵巧手星动 XHAND1 系列，已能完成拧螺丝、倒水等上百种复杂灵巧操作任务；最新发布的星动 L7 及 Q5 更是将这些能力向餐饮前厅与后厨协作深度拓展。宇树科技的 Unitree G1 量产版以极高的性价比打入市场，成功展示了开启核桃、厨房烹饪等精细技能，成为小微零售与餐饮场景的潜力机型。

家庭教育与陪伴场景中，机器人深度融合了知识传授与情感关怀功能。通过 AI 私教，机器人可实现作业智能批改与知识点拆解，手把手指导编程与艺术启蒙，并开展多语种情景对话训练。基于微表情与语音分析，机器人能实时识别情绪波动，生成焦虑疏导话术并提供共情互动（如故事讲述、定制化游戏）。此外，机器人可同步构建健康管理闭环，监测体与心率异常并触发服药提醒、送药喂水及紧急救援；更可通过模拟校园/家庭社交场景的角色扮演游戏，系统性培养儿童的沟通技巧与协作能力。相关应用正加速向消费与康养市场渗透。优必选的 Welli 优颐然机器人集成语音、触觉、视觉多模态交互，支持健康监测与情感陪伴，能通过情绪识别算法分析儿童面部表情和语音语调并动态调整互动内容；同时，其消费级 Alpha 系列全面升级了 AIGC 对话与编程教学，康养机器人也正逐步在高端养老社区落地。维他动力推出的智能伴随机器狗 Vbot 创新引入迪士尼动画法则设计交互动作，通过全模态交互精准捕捉用户情绪并提供心理疏导，其 VLA 模型统一了“看懂-听懂-做对-解释清楚”的全链路，目前已与京东达成战略合作进军消费市场。蔚蓝科技的 BabyAlpha 机器狗升级了 GPT 多模态交互，内置大语言模型并拓展了儿童绘本阅读等互动功能。傅利叶智能针对康养场景推出 GR-3，放弃了工业风硬朗外壳转而采用圆润温暖的软触感包覆材质，主打亲和力与情感交互，不仅能识别老人情绪进行聊天安抚，更致力于满足辅助陪护与长者看护的物理需求。

4.2 工业

工业具身智能（Industrial Embodied Intelligence）是指通过多模态感知融合、动态环境建模与自主决策闭环，使机器人等智能体在限定工业场景中，依据生产目标要求及工艺流程约束，完成生产作业任务的技术与系统。其核心特征表现为：基于多源异构数据的环境感知理解能力、面向产线动态调整的实时决策规划能力，以及适应工艺迭代的精准柔性执行能力，最终实现制造过程的智能化升级、质量稳定性跃升与安全

生产保障。

相较于开放环境下的通用具身智能，工业场景具有结构化程度高、工艺流程标准化、环境扰动可预测等显著特征，因此工业具身智能有望更早实现大范围落地和推广。然而，现代制造业呈现柔性化发展趋势，具有市场需求定制化、产品更新迭代快、工艺品质要求高等特点，决定了制造产线要在保证制造加工精度的同时，还须兼顾多款产品和兼容多种工艺。因此，柔性制造时代的工业具身智能面临两大独特挑战。

柔性适配与工艺精度的动态平衡：在应对多品种、小批量生产需求时，机器人既要保证制造精度（如汽车装配精度往往需要达到丝级 ±0.05mm），又要灵活应对因制造品类和工艺动态变化引起的工况变化、产线重构等挑战。此外，由于小批量制造难以通过制造规模来摊平成本，柔性制造产线往往无法做到像传统大规模制造产线的精度。例如，新能源汽车车型更新迭代快，往往需要一条产线混产多种车型，产线精度与传统燃油车制造的专用产线无法同日而语，但并不能因而降低工艺水平要求。因此，如何在低精度产线上完成高精度工艺，也为工业具身智能带来了巨大挑战。

通用技能与专门工艺的有机统一：智能制造机器人既要具备跨领域的基础操作能力，如抓取、放置、装配、拧紧、轨迹跟踪、曲面随形等，又要掌握面向特定制造工艺的专家级技能。例如，智能焊接机器人既需要具备焊缝跟踪的基础技能，又需要具备焊接工艺知识，如不同焊接工艺、母材种类、坡口形状等条件下的工艺参数设置，以及知识-数据双驱动的焊接工艺控制技能，通过对焊接熔池的实时监控和闭环反馈，对焊接电流、焊枪角度、移动速度等工艺参数进行实时调控，从而确保焊接质量。因此，如何基于基础通用操作技能实现特定复杂制造工艺，是工业具身智能面临的独特挑战。

工业具身智能的核心技术大致可分为工业之眼、工业之手和工业之脑，如图 4-1所示。工业之眼通过精密感知系统对制造对象和过程进行精准监测，需突破结构化场景的专用感知，实现面向复杂、可变工况的多模态、通用化感知与理解。需要注意的是，工业之眼是一个广义的概念，并非只包含视觉感知，也涵盖力触觉、超声波、电信号等多种模态的感知。工业之手是指工业机器人通过实时、精准操控来实现某种制造工艺的执行与调控，需突破面向预设工艺的离线编程限制，实现面向复杂制造工艺的自适应、智能化、在线化调控。工业之脑是更为宏观的规划和决策，一般是指对整个制造产线的智能调度、排产优化与最优控制，需突破固定制造流程的优化，实现面向多任务、多工段、多工位全局排产优化，同时还需要灵活适应生产任务的动

态变化（如订单变更、插单等）。当然，整个工业产线的智能、高效运转，是分布在各处的传感器（眼）和执行器（手），在集中或分布式部署的工业之脑的统一监控和调度之下协调运行的结果。

图 4-1　工业具身智能的挑战与核心技术

4.3　农业

随着全球农业面临劳动力短缺、资源有限和气候变化等多重挑战，农业机械化和智能化成为提升农业生产效率和可持续发展的关键路径。农机装备、农业机器人的具身智能技术，作为农业现代化的重要组成部分，正在迅速演进。在“感知-决策-控制（执行）”架构下，通过集成定位与导航技术、传感器技术、通信技术、具身智能与数据处理技术、感知-决策-控制一体化设计技术等具身智能（Embodied Intelligence）前沿技术，具身智能技术不仅显著提升了农业机械的自主作业能力，还实现了精准农业、资源优化配置和环境友好型生产，其典型硬件架构如图 4-2所示。

围绕农机装备的具身智能，当前实践主要集中在以下几个方面：

（1）自动驾驶系统

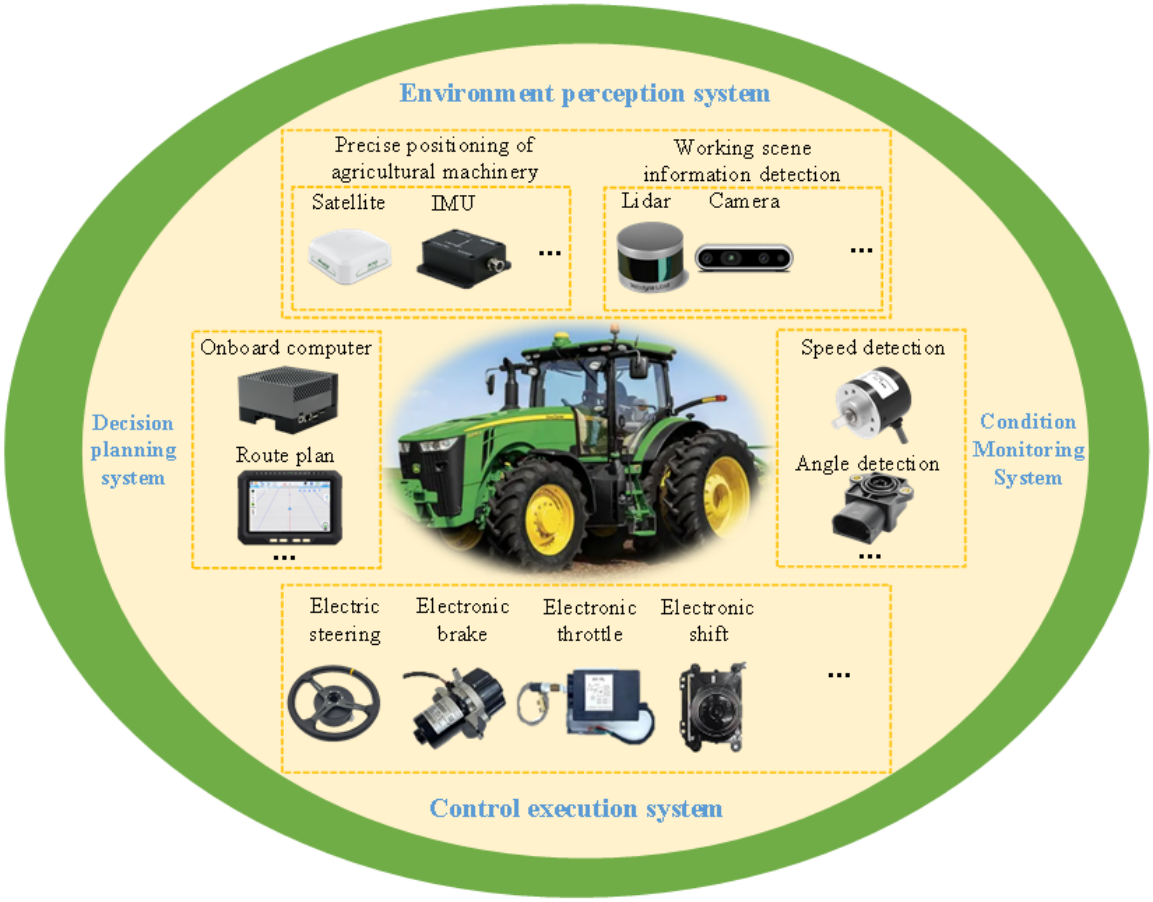

图 4-2　农机装备具身智能硬件架构

现今，全球主要的农机制造商，如约翰迪尔（John Deere）、CNH 工业（CNH Industrial）、潍柴雷沃、中国一拖、中联重科、极飞科技、华测导航等，都已经推出了有自动驾驶功能的农机设备。这些设备集成了高精度 GNSS、LiDAR 和摄像头等传感器，并搭载了先进的自主导航算法。它们能够在预设路径上自主作业，实现自动导航。此外，这些设备还能识别障碍物并调整路径。比如，约翰迪尔 2025 年推出 Next-Gen Autonomy Kit 与 AutoTrac 2.0，8R/9RX 系列自动驾驶拖拉机借助搭载 16 路立体视觉与 Nvidia AI 芯片，360° 感知、24 米外避障、±2.5cm 定位，能在复杂农田环境中实现自主导航，支持高效无人作业。潍柴雷沃与华为合作，推出 CVT 智能拖拉机（340 马力），通过北斗 +5G+ 边缘计算，适配大田与丘陵，完全无人驾驶。华测导航 2025 年发布 NX612 自动转向系统，采用多模 GNSS（RTK/PPP），导航误差在 ±2cm 以内，适配全地形。

（2）农业机器人集群

集群智能技术作为农业无人机（UAV）与农业无人车（UGV）执行空地协调任务的基础性支撑。针对 UAV 与 UGV 协作，根据编队元素组成的不同，其协作组织形式主要为多 UAV 空中协作、多 UGV 地面协作及 UAV-UGV 空地协作。协作编队的组织形式根据各节点信息交互方式不同，一般分为分布式、集中式和分散式，由于分布式策略是基于局部信息交互完成协作编队，其与生物集群的信息传递形式更贴近，同时可以兼顾作业精度与信息交互负担等多方面性能，分布协作具有更为突出的应用优势与探索价值。而在局部信息交互下，实现集群协作作业，高效精准完成农业任务，成为农业 UAV-UGV 协调任务的关键环节。集群协作在农业领域主要执行遥感、测绘、病虫害及杂草监视、施药等农业植保任务，以及农业收获、运输任务。比如，CNH 工业的自动驾驶系统支持机群协同、变量作业，在北美、南美大规模应用。中国一拖的东方红 LF2204/LF1104 无人拖拉机，通过北斗 RTK 的导航误差控制在 ±2.5cm 以内，支持自动掉头、集群作业。

（3）精准农业应用

具身智能技术在精准农业里的应用愈发广泛，如自动播种、精准施肥、精准喷药、智能收割等环节都包含在内。智能农机能实时监测土壤湿度、养分含量以及作物生长状态，依据不同区域的需求自动调整作业参数，精准利用资源。比如，施肥无人机利用传感器数据实时对施肥量和施肥区域加以调整，防止因过量施肥而产生环境污染和带来经济损失。中联重科的水稻智能插秧机搭载智驾系统，自动驾驶、一人多

机、效率翻倍，其无人收割机经过万亩测试。极飞科技的 APC2 自驾仪，导航误差在±2.5cm 以内，适配拖拉机、插秧机、采棉机，App 一键规划路径。

（4）智能避障与安全系统

智能避障系统借助多传感器融合技术，能实时对周边环境进行感知，识别潜在障碍物，自动对作业路径加以调整，从而保障农机在复杂农田环境里安全运行。系统借助 AI 算法可预测障碍物动态变化，从而提前作出反应。另外，智能安全系统也有紧急停止、故障检测、远程监控、安全警报等功能，这会进一步提高农机的安全性与可靠性。比如，CNH 工业的智能避障系统会实时对 LiDAR 和摄像头的数据进行分析，从而自动避开障碍物，保证作业时的安全性。

（5）农机健康管理与预测性维护

智能农机借助物联网与大数据分析，能实时监测机械部件的状态，开展健康管理，还能进行预测性维护。系统分析传感器数据，可预测潜在故障，提前维护设备，减少设备停机时间与维修成本。比如，通过振动传感器和温度传感器的数据进行分析，就能及时察觉发动机或者传动系统的异常，从而进行预防性维护，延长设备的使用寿命。约翰迪尔通过 Operations Center 实现远程机群健康管理和维护。

（6）农机能源管理与绿色农业

绿色农业在未来农业发展中是重要方向，车载具身智能技术在能源管理方面起着关键作用。发展像电动或者混合动力无人农机这类高效的节能动力系统，既减少了对化石燃料的依赖，又减少了碳排放与环境污染。智能能源管理系统可实时监控农机的能源使用情况，并进行优化，从而提高能源利用效率，促进农业设备向绿色方向转变。比如，无人拖拉机有电动动力系统和智能能源管理系统，可以优化能源分配保障作业时间长且高效。

无人驾驶农机要实现自主作业，精准定位与导航是基础，这样农机才能按预定路径高效、安全地完成各种农业任务。相关关键技术包括：

（1）全球导航卫星系统（GNSS）

多系统融合定位：除了传统的 GPS、GLONASS、Galileo 和北斗系统，现代农机常采用多系统融合技术，通过多卫星系统的信号（GPS、GLONASS、Galileo 和北斗等多个卫星系统的信号）可以提高在复杂农田环境中的定位精度和可靠性。实时动态定位（RTK）与网络 RTK：RTK 技术利用基准站发出的差分修正信息，实时提升

GNSS 定位精度至厘米级。最新的网络 RTK 服务借助广域网络，连接多个基准站，实现了更广泛的高精度定位，增强了定位的灵活性与覆盖区域；增强型 GNSS 信号处理：将多路径抑制和抗干扰技术等先进信号处理算法相结合，以此提高 GNSS 信号的稳定性和准确性。运用机器学习算法来优化信号解算流程，增强在复杂环境里的定位性能。

（2）惯性导航系统（INS）

高精度 MEMS 惯性传感器：最新的 MEMS 惯性传感器精度更高，稳定性也更强，在微小的振动与冲击环境下，能给出可靠的定位数据。随着先进的制造工艺和材料科学的发展，INS 系统也更加紧凑、轻量化，使其能够集成到各类农机装备中。GNSS/INS 深度融合技术：利用深度学习和自适应滤波算法，能够明显提高整体定位系统的精度和鲁棒性。GNSS 能够提供长期的绝对定位，而 INS 能提供短期的相对定位。二者相结合，就能达成高频率且延迟低的定位更新，这样农机在复杂地形与恶劣天气条件下的持续导航能力就能得到保障。

（3）视觉导航与里程计

先进的视觉同步定位与地图构建（Visual SLAM）：将深度学习模型和实时图像处理技术相结合，提高 Visual SLAM 在动态和未知环境里的稳定性与准确性。采用多摄像头系统结合高分辨率雷达传感器技术，生成精细化的环境地图，以实现高精度的自主导航功能。深度摄像头与立体视觉：采用具备高动态范围（HDR）且针对低照度环境优化的深度摄像头，提升在不同光照环境下的三维环境感知能力。把立体视觉算法和深度学习相结合，提高障碍物检测与路径规划的准确性与实时性。多模态传感融合：将视觉信息跟雷达、超声波之类的其他传感器数据进行多模态融合，让环境感知的全面性与准确性都得到提升。比如说，将视觉数据和毫米波雷达数据相结合，能实现更可靠的障碍物识别与避障功能。

（4）高精度数字地图

高精度数字农田地图：借助高分辨率遥感数据、无人机航拍资料及地面传感器信息，构建高精度、动态更新的农田数字地图。集成土壤类型、作物分布、水源位置以及灌溉系统等多维度信息，让农机能精准作业，实现智能决策。智能路径规划算法：采用基于深度强化学习和多目标优化的路径规划算法，动态地对作业路径加以优化。算法能够即时适应农田环境的变化，避开障碍物，调整作业范围，并优化资源配置，达成多目标优化效果，从而提升作业效率及资源利用率。自适应地图更新机制：将实

时传感器数据跟无人机实时监测相结合，让农田地图得以自动更新与维护。借助云端计算平台，能实时同步、共享地图数据，以支持多台农机协同作业，还能对其动态路径作出调整。

在感知层面，农机通过多种传感器协同工作获取环境与作业状态信息，典型配置包括：

（1）激光雷达：高密度激光扫描仪可生成精确的三维环境模型，以识别农田里的障碍物、地形变化和作物分布。固态 LiDAR 技术体积小、成本低、可靠性高，所以逐渐在农业无人机和农机装备里应用起来，提升了环境感知的效率和精度。

（2）雷达传感器：在恶劣天气下，毫米波雷达与超声波雷达仍能稳定工作，可用来检测远距离障碍物以及动态环境变化。多频雷达技术不断发展，检测精度与抗干扰能力都提高了，雷达传感器在复杂农田环境中的应用也变得更广泛了。

（3）高分辨率摄像头：拥有多光谱与高光谱摄像头，可捕捉作物生长状况、病虫害信息以及土壤条件。将图像处理算法和深度学习模型相结合，达成作物分类、健康监测以及精准施肥的目的。例如，多光谱摄像头可以检测作物的叶绿素水平，从而指导精确施肥与灌溉。

（4）作业监测传感器：土壤湿度传感器借助电容、电阻或者时间域反射测量（TDR）技术，对土壤水分含量进行实时监测，以实现精准灌溉与水资源管理。高精度土壤湿度传感器可依据不同地块的湿度情况，自动调控灌溉系统，以优化水资源利用。

（5）作物生长监测传感器：作物生长监测传感器包含多光谱成像、叶面积指数（LAI）传感器以及生物光谱传感器等，这些传感器可用于实时采集作物生长数据，从而指导精准施肥、喷药与收割工作。借助传感器网络，农机可动态调整作业参数，从而全面监控与管理作物生长。

（6）机械作业传感器：机械作业传感器包含转矩传感器、振动传感器与温度传感器，其目的是监测农机作业状态和机械健康，防止故障，优化作业参数。实时采集与分析数据，为预测性维护提供支持，从而减少设备停机时长和维修成本。。

（7）农机装备数字孪生：构建农机设备的数字孪生模型，借助实时传感器数据与物理设备同步，达成设备状态的虚拟监控以及故障预测。数字孪生技术可模拟设备在不同作业环境下的性能表现，进而优化操作参数，提高作业效率，并延长设备使用寿命。

在完善本体感知能力的基础上，还需要稳定可靠的通信与网络基础设施来支撑多机协同与云端智能，主要技术包括：

（1）车联网（V2X）：V2X 技术可以实现农机之间以及农机与农田基础设施的实时信息交换，包括车与车（V2V）、车与基础设施（V2I）、车与云端（V2C）等多种通信模式。V2X 能让农机协同作业，共享作业路径与任务分配，提升整体作业效率。例如，多台无人农机通过 V2X 技术可以协调各自的作业区域，避免重复工作，从而实现大规模农田的高效管理。

（2）5G 通信：5G 网络以其高速率、低延迟和大连接的特点，支持大规模数据的实时传输和处理。5G 的边缘计算能力能让无人农机在本地进行复杂数据处理与决策，可以减少对云端的依赖，还能提升响应速度和系统稳定性。5G 网络能够支持无人农机进行实时视频传输与远程操控，从而实现精准作业及远程监控。

（3）低功耗广域网（LPWAN）：如 LoRa 和 NB-IoT 等低功耗广域技术，适合在覆盖广、能耗低的农业环境中承载大规模传感器网络的数据传输。LPWAN 技术在农业物联网中的应用，使农机能够高效采集与传输环境数据，助力精准农业得以施行。

（4）安全通信与数据加密：保障农机与云端数据的传输安全，避免数据泄露、篡改，保证数据安全与隐私保护。安全协议和身份认证：通过使用先进的安全协议及多层次的身份验证方法，阻止非法访问与网络攻击，确保通信系统的安全。

在通信与算力基础设施之上，具身智能与数据处理技术为农业具身智能提供决策与优化能力，典型方向包括：

（1）机器学习与深度学习：机器学习和深度学习算法在图像识别、路径规划及决策制定中扮演了重要角色。卷积神经网络（CNN）在作物识别与病虫害检测方面有应用，强化学习算法则用来优化作业路径并做出策略决策。比如，深度学习构建的图像识别系统，可以精准识别不同作物的生长阶段与病虫害情况，做到精准施肥和喷药。

（2）边缘计算与云计算：边缘计算这一技术能让数据在本地设备上实时处理，减少数据传输的延迟，提高系统响应速度。农机能在本地对传感器数据予以处理，快速作出作业决策，保障作业高效且安全。云计算平台主要用于大规模数据存储、数据分析及模型训练，支撑无人农机的持续性能优化与智能化升级。

（3）大数据与农业信息系统：借助收集和分析农机作业数据、环境数据以及作物生长数据，打造农业大数据平台。将地理信息系统（GIS）与遥感技术相结合，达成精准的农业管理，给智能决策提供支持。通过大数据分析，可以预测作物产量及病虫

害风险，从而指导农机进行精准作业。

（4）具身智能驱动的自适应作业控制：在此基础上，引入具身智能系统，将感知、认知与动作控制闭环打通，使农机能够根据实时环境状态和任务需求动态调整作业策略与机械行为。通过多模态感知融合技术综合利用视觉、雷达和触觉等多源数据，系统可以在复杂农田环境中形成更加全面的场景理解，并结合强化学习、模仿学习等方法持续优化策略，从而提升农机的自主操作能力、作业效率和作业质量。

（5）农业机器人感知-决策-控制一体化设计技术：机器人传统设计中感知、决策与控制的割裂，已成为提升机器人环境适应性、任务鲁棒性与作业效能的瓶颈。机器人一体化具身设计强调从系统整体出发，实现感知信息流、决策逻辑与控制执行的内在耦合与协同优化，是突破当前技术瓶颈、实现农业机器人真正自主与智能化的关键路径。面向未来无人农场与精准农业的发展需求，如何实现农业机器人感知、决策与控制功能的深度耦合与一体化设计，已成为突破现有系统性能瓶颈、提升作业效率与适应性的关键科学问题与工程挑战，直接决定农业装备的实际作业效能。需要研究一体化设计理论与系统设计架构，包括分层一体化设计架构，感知-决策-控制紧耦合的端到端一体化设计架构，轻量化、低延迟、高可靠的一体化控制流机制，机器人集群多智能体一体化设计架构与通信决策控制融合机制，感知-决策-控制全链路的联合优化与学习算法，资源约束下的一体化轻量化设计与性能权衡，农业场景中智能行为的系统级涌现，端到端的一体化建模与控制，以及视觉-语言-动作大模型与农艺知识融合。

农业具身智能正从“自动化执行”向“认知型自主”转变，物理世界基础大模型与感知、决策、控制深度融合，从单一专用农机向开放环境、多机协同、自主进化的通用智能体发展。以上技术将共同支撑农机装备、农业机器人从局部自动化走向全程自主化与智能化。未来，农机装备、农业机器人将进化为具备群体智能、能自我适应的具身智能体，真正融入复杂的农业生态系统。

4.4 交通

当前，随着城市化进程的加速推进以及出行需求的持续攀升，交通管理正遭遇前所未有的挑战。一方面，道路拥堵、交通事故频发等问题亟待解决；另一方面，行业对精细化管理与智慧化升级的需求愈发迫切。具身智能技术凭借其独特的多模态感知、自主决策及精准执行能力，为破解交通治理难题提供了新路径。通过将感知设备

与执行单元深度融合，该技术不仅可以有效提升交通系统的风险预警与应急处置能力，更推动了交通管理从被动响应向主动干预的模式转变，其在道路基础设施安全、自动驾驶与智能物流等领域的典型应用如图 4-3所示。

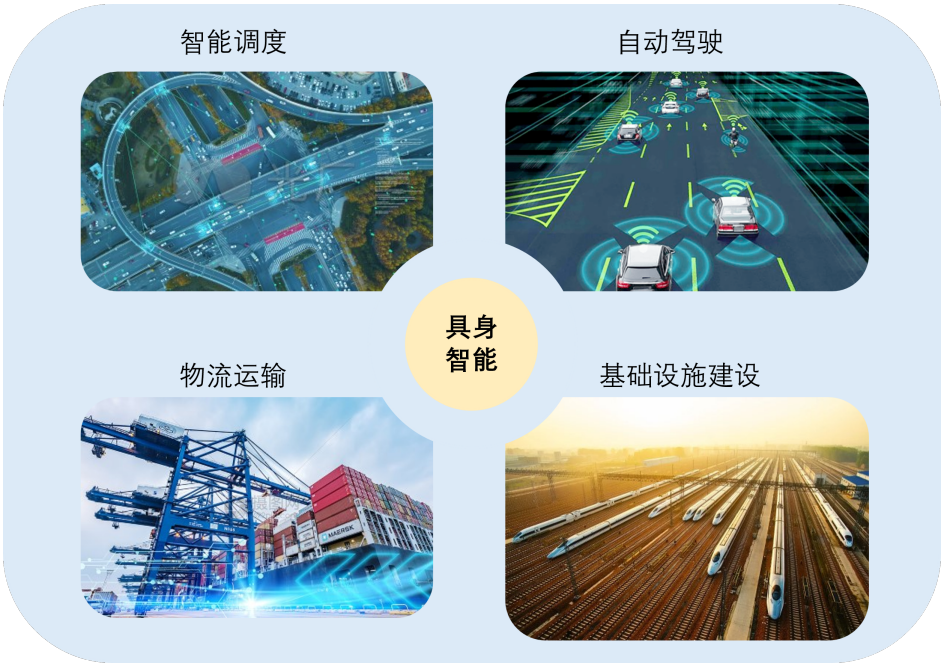

图 4-3　具身智能在交通各领域的应用

（1）基础设施安全领域

基础设施安全主要涉及安全检测，长期以来，基础设施安全检测主要依靠人工巡检和单一传感器技术进行监测，不仅响应速度慢、误报率高，而且还难以适应复杂环境，相较于传统检测方式，具身智能系统具备更强的环境适应能力，能够自主学习优化，从而提升监测的可靠性和效率。以铁路安全检测为例，华为开发的智能监测系统融合毫米波雷达、振动光纤和 AI 视觉技术，通过多源数据融合算法，不仅实时识别轨道入侵事件，还将误报率降至万分之一以下。同时，借助联邦学习技术，不同区域的监测系统能够协同优化，进一步提升整体性能。同样具有代表性的港珠澳大桥的健康监测系统，则依托 2000 多个传感器，构建一个实时联动的智能网络，在台风来临时能够实现从数据采集到预警发布的毫秒级响应。

这些系统依托具身智能的感知-决策-执行闭环架构，将传感器数据（如雷达点云、触觉振动、视觉图像）与物理环境实时交互，结合大模型的语义理解能力，不仅提升了全天候检测可靠性，还通过端云协同实现数据驱动的持续进化，为交通基础设施安全运维提供了高效、低成本的智能化解决方案。

（2）自动驾驶领域

目前，具身智能正推动自动驾驶从模块化设计向端到端架构转变和过渡，这种新方法能直接将路况感知转化为驾驶决策，大幅提升系统响应速度和复杂场景处理能力。相较于传统的模块化设计，端到端系统可以直接从海量真实驾驶数据中学习更接近人类司机的决策逻辑而不是基于预先设定的规则定义，最重要的是，这种具身架构可随着数据积累不断提升在极端场景下的表现。目前，特斯拉的 FSD V12 系统已经通过纯神经网络实现了从摄像头输入到车辆控制的端到端决策，大幅提升了复杂路况下的拟人化驾驶表现，而国内小鹏汽车的 XNGP 系统三模块端到端架构也已经量产上车，支持无图城区导航辅助驾驶，并通过联邦学习实现跨区域模型优化。2025 年下半年至 2026 年初，端到端大模型进一步成为自动驾驶行业的重要技术共识，L3/L4 路线与商业化落地同步提速。小鹏推出第二代 VLA 模型，实现从视觉信号到动作指令的端到端生成；华为宣布 ADS 4.0 将于 2026 年面向高速 L3 商用。

相比传统模块化架构需要经过感知、决策、规划等多个独立环节的处理流程，端到端自动驾驶技术实现了从原始感知数据到控制指令的直接映射，不仅大幅提升了系统响应速度，更让自动驾驶车在面对突发状况时，展现出接近人类的本能反应的应变能力。

（3）物流运输领域

物流运输领域，具身智能有望降低流通成本，成为形成高效、快捷、智能化的物流体系的关键因素。当前物流领域包括拣选机器人、叉取机器人、搬运机器人、料箱机器人等。具身智能技术的赋能，可以在仓储、装卸、搬运、分拣、包装、配送等环节提升工作效率和管理水平。物流机器人将更加智能化，具备更强的自主决策和学习能力，能够适应更复杂、多样化的任务，不仅局限于传统的仓储和物流行业还将渗透到制造业、农业、医疗、教育等领域，提高各行各业的智能化水平和生产效率。例如，亚马逊近期在其仓库运营中，已经在测试由其投资的公司 Agility Robotics 开发的人形双足机器人 Digit，综合全面完成主要包括卸载货车、搬运箱子、管理货架等任务，大幅提高了仓库作业的效率。星动纪元联合北自科技推出“具身智能物流仓储解决方案”，将双足人形机器人星动 L7 及具身大脑 ERA-42 应用于“货到人”拣选环节，实现端到端 VLA 模型在物流仓储场景中的真实落地；德马科技与智元机器人围绕物流场景建设“人形机器人数据采集工厂”。

（4）智能调度与运维领域

交通枢纽的运维场景下，不同于传统响应滞后的静态信号配时，具身智能系统能够像经验丰富的交警一样，根据实时车流动态调整相位周期——当检测到公交专用道有车辆接近时，可提前触发绿灯响应；发现行人闯红灯行为时，又能联动周边摄像头进行多角度复核。杭州亚运村周边的路侧单元更是展现出类人的预判能力，在暴雨导致能见度下降时，自动切换为基于毫米波雷达的冗余感知模式，确保异常事件的识别准确率不因天气影响而衰减。这些实践表明，具身智能正在将离散的交通要素转化为有机协同的智能体网络，通过对环境的持续学习，展现出应对复杂动态环境的独特优势。

4.5 能源与电力

随着具身智能和大模型技术的发展，学术界与工业界已经在电力系统中开展了一系列机器人带电作业与智能巡检的关键技术研发和示范应用，从理论方法、感知与控制算法到专用装备研制持续演进，并逐步在输电通道、新能源场站、变电站与储充设施等关键环节展开部署，正形成“空地协同、多机协作”的智能运维体系，为后续大规模落地奠定了基础。

（1）输电线路巡检与通道运维。传统输电通道巡检依赖人工攀爬铁塔或沿山路步巡，不仅效率低、危险系数高，而且难以及时发现隐蔽缺陷。具身智能驱动的巡线无人机通过搭载多模态传感器（可见光、红外、激光雷达等），结合自主航线规划与目标识别能力，可在复杂山地、跨江跨峡等场景下实现自动起降、自动巡检和缺陷告警。一些线路场景还引入了轨道式、轮爬式或多臂攀爬机器人，在导线、地线或塔身上进行近距离“贴身”巡检，对异物、散股、腐蚀、金具松动等风险进行精细化检测，并与地空无人机协同完成大范围巡查与重点部位复核。例如，联想集团联手复旦大学开展的电力巡检具身智能项目，利用 VLA 大模型进行任务理解和操作决策，并结合开放词汇 (open-vocabulary) 目标搜索与定位技术，在大负载六足机器人平台上完成了复杂地形巡检、搜救和勘探等多场景示范应用，展示了具身智能在电力巡检场景中的可行性与扩展潜力。

（2）新能源场站运维与设备清洁。大规模光伏电站、风电场的运维高度依赖高频巡检与组件清洁，而人工巡检和清洗在高温、高海拔和沙尘环境下面临显著挑战。具身智能机器人在这一场景中已经出现多种落地形态：一类是面向光伏电站的自动清扫与清洗机器人，可在组件阵列表面自主规划路径、跨越排间间隙，并根据灰尘堆积

情况动态调整清洁策略，显著提升发电效率；另一类是塔筒/叶片检测无人机与爬壁机器人，能够在强风、高空环境下获取高分辨率图像和结构振动数据，结合智能分析模型完成裂纹、掉漆、结冰等缺陷识别，为风机健康管理提供高频、低成本的数据支撑。

（3）变电站与配电设施的智能巡检与操作。变电站传统巡检多由值班人员定时巡视或依赖固定摄像头远程监控，难以及时、细粒度地感知设备状态。具身智能巡检机器人通过自主导航在站内巡游，完成红外测温、局放监听、仪表读数识别、异物入侵检测等任务；部分场景中，双臂或多臂操作机器人已经开始承担开关倒闸、刀闸分合、避雷器更换等高危操作，实现“人从高危一线撤出、机器人在前线执行”的运维模式。在城市配电网场景中，轮式或履带式配电房巡检机器人也已用于电缆头测温、局放检测和环境监测，与远程监控平台共同构成无人值守配电房的关键基础设施。在配网带电作业领域，南方电网广东广州供电局自主研制的混合现实（MR）遥操作带电作业机器人“悟空”，以及国网天津市电力公司研制的配网自主带电作业机器人，分别通过 MR 交互和双机械臂协同作业模式，完成了引线搭接和 10kV 带电接引线等现场试验，为高风险工况下的人机协同作业提供了重要实践依据。不过，这类系统当前在作业灵活性、自主决策以及对环境瞬时变化的自适应方面仍然在很大程度上依赖人工操作，后续有待通过具身大模型、强化学习等技术进一步提升带电检修作业的自主化水平与整体质效。

（4）储能、电动车充换电等新业态中的具身协同。随着分布式光伏、户用储能和电动汽车的普及，分布式能源设施数量和空间分布复杂度急剧上升。具身智能机器人可以承担大型电化学储能站的电池舱巡检、泄漏检测和紧急处置任务，也可以在城市公共充电站或换电站中执行电缆自动插拔、电池包搬运与仓储管理，实现“无人看守、按需响应”的柔性服务模式。结合车端的具身智能系统，还可以实现车-桩-站之间的协同调度，例如在低谷时段引导自动驾驶车辆自主前往指定站点充电或换电，在突发停电、极端天气等场景下实现具身群体应急保供。

总体来看，具身智能在能源与电力行业的价值，不仅体现在替代高危、高空、高压环境下的人工作业，更在于通过连续、多模态的实时感知和闭环决策，构建起“感知—理解—执行—评估”的全生命周期智能运维体系。未来，随着大模型与行业知识图谱的深度融合，电力具身智能有望从“场景化单点应用”走向“跨场站、跨电压等级的协同调度与自主运维”，在保障能源安全、提升系统韧性和支撑“双碳”目标方面发挥更大作用。

第五章　具身智能未来发展趋势

具身智能的研究正朝着应用多样化和能力深入化方向不断发展，未来的核心将聚焦于提升跨场景通用性、自主学习和环境适应能力、人机交互的效率，以及多模态感知与决策的融合水平。从研究广度来看，具身智能的应用正逐步超越传统机器人学的范畴，深入渗透到智能医疗、自动化制造、精准农业、智能交通等关键领域。例如，近期在手术辅助机器人与智能康复系统方面取得的进展，不仅提升了机器人在复杂手术中的操作精度，也增强了患者在康复过程中的个性化适应能力。在自动化制造中，具身智能正向多任务协同方向发展，结合强化学习与大规模预训练模型，使工业机器人能够在非结构化环境中自主学习新技能，从而优化流程并提升效率。从研究深度来看，未来的具身智能系统将更加强调自主学习和环境适应能力，这对于通用智能体的发展具有重要推动作用。随着大规模世界模型与先进交互决策算法的快速演进，具身智能在跨环境适应与动态规划方面的能力正在持续提升，初步展现了较强的泛化和自主决策水平。

面向未来，具身智能研究将呈现以下四大趋势：（一）从单一模态向多模态感知与交互的闭环机制发展，强调感知与行动之间的紧密耦合；（二）从静态场景理解向动态环境预测与适应演进，关注世界模型的实时构建和推理能力；（三）从限定场景训练向开放环境迁移，探索在不确定和新颖环境中实现稳定的迁移学习与自适应决策；（四）从单一智能体向多智能体协作拓展，研究群体智能系统中的协同机制和涌现能力。这些发展将共同推动具身智能从实验室走向真实世界应用，为解决复杂环境中的感知与交互问题提供新范式。

5.1　具身智能关键技术发展趋势

随着深度学习和大模型技术的快速发展，具身智能有望在多模态感知与认知融合、自主决策与闭环控制、自适应学习与知识进化、仿生形态设计与运动控制、群体协同与分布式智能、安全可信与伦理治理等关键技术方面取得重要进展。

在感知层面，具身智能系统将视觉、触觉、听觉等多源异构传感器数据进行深度融合，使得其具备接近人类水平的综合环境理解能力。例如，基于高精度的仿生传感器网络，智能体能够具备对环境物理特性（如压力、温湿度、材质纹理）的精确解析

能力；借助多光谱成像和量子传感技术，智能体的视觉系统可获取更广泛的光谱信息，实现红外、紫外、毫米波的全频段覆盖，超越可见光限制；通过柔性电子皮肤结合神经形态计算技术，实现对物体形变与弹性参数的高灵敏度实时感知。在认知层面，具身智能系统将结合实时感知信息与先验知识，动态构建和维护环境的内部表征模型，为复杂场景下的推理与决策提供支持。其中，基于神经符号推理的混合认知架构，有望成为主流方法，通过大模型驱动的跨模态对齐技术，提升智能体的语义理解与场景预测能力。并基于 4D 时空记忆机制，增强智能体对动态场景的时序理解和反事实推理能力。未来研究方向主要包括仿生传感器的小型化与集成、多模态数据融合架构设计，以及基于大模型的因果推理能力优化等。

在自主决策方面，具身智能系统将结合大模型所具备的强大先验知识与小样本学习能力，减少对大规模标注数据的依赖，使其更快速地适应新环境与新任务。基于强化学习与大模型的融合，具身智能体的环境建模、预测与推理能力将得到显著增强，有助于减少物理试错的风险与成本，提升智能体的决策效率与安全性。通过引入分层强化学习架构，具身系统可在高层任务规划与低层动作控制之间建立有效连接，支持长时间、复杂任务的稳定执行和灵活调整，从而提升系统的适应性与可扩展性。在闭环控制方面，具身系统将广泛应用自适应控制与智能优化算法，实现控制参数的动态更新，以应对复杂、非线性和变化多端的环境条件。高精度、高带宽的感知反馈系统是实现精细控制的基础，能够保障运动规划的准确执行与整体系统的稳定运行。借助实时数字孪生技术，具身系统可在虚拟环境中进行快速仿真、测试与策略优化，从而提升响应速度与故障预判能力。智能优化算法与鲁棒反馈机制的深度融合，也将增强具身系统在干扰环境下的自我调节与修正能力，有助于提高任务完成率与系统长期运行的稳定性。未来的研究方向可能包括神经拟态控制、量子强化学习的交叉探索、分布式智能控制理论的深化，以及面向因果结构的控制建模等。

在学习能力方面，具身智能系统正逐步摆脱传统的离线训练范式，转向结合自监督、持续学习等方法，使系统能够根据环境与内部状态的持续变化，自主调整其感知、决策与行为策略，更有效地适应新任务与未知挑战。增量学习和在线学习的应用，使得系统可以在任务执行过程中不断更新其认知模型和行为策略，避免灾难性遗忘，不再依赖于完整的重新训练过程，从而提升在动态环境中的持续应变能力。元学习的引入将使系统能够“学会如何学习”，即便在新场景下，也能快速迁移过往经验，提升适应效率和泛化性能。结合多任务学习和强化学习，系统可在复杂现实场景中并行处理多种任务，并优化整体策略，实现更高的准确性与资源利用效率。在知识

进化方面，具身智能将探索多智能体间的知识共享与协同进化机制。借助分布式交互和信息共享，系统能够不断积累、整合并提升知识水平。通过神经进化和遗传算法等技术，系统可模拟生物演化过程，逐步优化神经网络结构与策略，实现自主的能力提升。动态知识图谱和自组织学习机制将支持多源异构知识的整合与迁移，增强系统在复杂情境下的适应性与创新能力。未来研究方向包括受量子机制启发的强化学习框架、生物遗传机制模拟的知识表示方法、基于拓扑分析的学习网络结构，以及去中心化的知识协作平台等。这些技术有望推动智能体从学习既有知识，迈向主动探索未知领域。

在形态设计方面，具身智能正突破传统刚性结构的局限，结合仿生原理与柔性材料、智能材料等新兴技术，推动形态的多样化、功能灵活性与结构可重构性。仿生关节与肌肉-骨骼协同驱动结构将提升运动的灵活性与敏捷性。采用折纸结构与软体材料的微型机器人，能够适应复杂或受限环境，在微观尺度上执行精密操作，拓展了应用边界。可变刚度执行器和 4D 打印材料的应用，将使系统根据任务需求主动调整自身结构与力学特性，更好地适应环境变化。在运动控制方面，通过结合高保真仿真与从虚拟到现实的迁移学习策略，智能体在复杂地形（如崎岖地面、楼梯）上的自主行走与操作能力将进一步提升。运动规划与轨迹优化技术的发展，将增强运动的平稳性、效率与稳定性。可变构型机器人通过集成拓扑优化与智能材料技术，实现多种运动方式的自由切换，如轮式、足式或软体爬行。基于介电弹性体和液晶弹性体的驱动机制，也有望实现微尺度下更高精度、更高柔顺性的交互与控制。未来，形态、功能与控制的协同设计将成为开发高性能智能体的核心方法。研究重点可能包括具备自修复能力的材料系统、基于神经形态计算的分布式感知与控制框架，以及能量效率更高的驱动与供能系统等，推动具身智能向更高水平的“类生命系统”演进。

群体协同能力是具身智能在多主体复杂交互场景中实现高效协作的关键，提升群体智能系统的协作效率与自组织能力，将有助于完成大规模、高复杂度的任务，并增强对动态环境的适应能力。通过分层协同架构、分布式共识算法与联邦学习等技术，多智能体可实现任务动态分解、最优分配及协作策略的在线优化。基于此，无人机编队、自动驾驶车队、人形机器人群体等异构智能系统的高效协同机制成为重要研究方向，目标是减少决策冲突，提高整体系统的协同性与稳定性。同时，量子通信等新型技术也可增强协同效率与安全性，拓展智能群体在极端或对抗性环境中的应用潜力。分布式智能强调智能体之间的物理交互、信息共享和联合决策，以实现更高水平的系统智能。通过端-边-云协同架构，可构建包含分布式推理节点与协调节点的

智能网络，支持大规模智能体集群进行任务分配、路径规划、冲突管理与动态协同。基于超大规模群体智能，尤其结合类脑计算架构如脉冲神经网络，将有望实现大规模并行学习、协作决策与复杂行为的群体涌现。未来研究方向包括异构智能体通信协议的标准化、群体行为的可解释性与可控性建模，以及在资源受限条件下的全局优化算法设计等。

安全可信的目标是确保具身智能系统在动态、多变、人机共融的实际环境中稳定、可靠且可预测地运行。从系统全生命周期来看，安全机制需贯穿硬件、算法、系统与应用层级。硬件层面通过冗余设计、故障诊断与安全隔离保障基本的功能安全与容错能力；算法层面需应对对抗性攻击、提升因果可解释性与透明性，增强决策过程的鲁棒性；系统层面则需要建立风险评估、运行时监控与应急处理机制，以实现全过程的风险管理与控制。在伦理治理方面，需要建立由技术开发者、使用者与监管者共同参与的协同治理机制。主要问题包括人机职责界定、数据隐私保护与安全传输标准的建立，以及确保智能体行为符合伦理规范与法律要求。构建基于概念激活向量的决策追溯系统，提升具身智能体决策的透明度与可信度；设计基于义务论与功利主义的混合伦理模型，构建形式化的道德推理机制，辅助具身智能体在复杂伦理情境中作出权衡与判断；研究物理不可克隆功能、光学神经网络加密等新技术，为底层系统提供更高等级的防护能力。同时，设置极端情境下的自毁机制，提升系统的整体安全性与抗攻击能力。未来技术突破点包括实时异常行为监测、形式化合规验证技术的自动生成、人类价值观嵌入的奖励函数设计等。

5.2 具身智能技术应用发展展望

5.2.1 从 VLA 到 WAM：世界模型驱动的范式跃迁

具身智能的算法架构正经历从视觉-语言-动作（Vision-Language-Action, VLA）模型向世界-动作模型（World-Action Model, WAM）的范式跃迁。2026 年初以来，这一趋势愈发显著：虽然 VLA 模型仍是当前主流技术路线，以 OpenPI、OpenVLA 为代表的系列工作表现稳定，但其固有架构在动态环境适应与长程规划方面已暴露出结构性瓶颈。世界模型通过构建可交互的物理环境仿真器，使智能体具备预测未来状态、评估动作后果的能力，为具身智能的认知升级提供了核心技术支撑。

国际前沿领域，Google DeepMind 于 2025 年末发布的 Genie 3 标志着生成式世界模型的成熟——该模型以 24fps 实时生成交互式三维环境，无需显式物理引擎即可

从数据中习得物理规律。NVIDIA Cosmos-Predict2.5 平台进一步统一了文本、图像到世界的生成范式，为机器人策略评估与闭环仿真提供了基础设施；2026 年初发布的 NVIDIA Cosmos Policy，进一步验证了 WAM 范式替代传统 VLA 模型的技术可行性，推动 WAM 逐步成为学术界与产业界的共识性技术路线。

国内方面，北京大学、智元机器人等科研机构与企业同步实现了 WAM 领域的关键技术突破，形成了与国际前沿并行发展的技术体系，为 WAM 范式的本土化落地与开源生态构建奠定了基础。北京大学团队联合北京人形机器人创新中心、香港科技大学发布的 WoW 世界模型，针对传统生成式模型只重视觉逼真度、不理解底层物理规律的缺陷，构建了物理规则约束下的交互式环境生成框架，实现了视觉真实度与物理因果准确性的联合优化，为 WAM 模型提供了高保真的环境仿真底座。2026 年 3 月，智元机器人发布动作序列驱动的具身世界模型框架 EVAC（EnerVerse-AC），同步开源全球首个具身世界模型专用评测基准 EWMBench，构建了覆盖数据增广、模型训练至性能评测的全链路技术闭环。该框架针对真机验证成本高、仿真与现实存在域偏差的核心技术瓶颈，基于世界模型实现训练数据的高效扩增，其仿真评测结果与真机实测结果具备高度一致性，可有效提升策略模型的筛选效率与任务成功率，为 WAM 范式与真机落地的深度融合提供了标准化开源基础设施。此外，清华大学联合研发的 Ctrl-World、星海图发布的 Fast-WAM、极佳视界推出的 GigaWorld-1 等成果，在世界模型的生成精度、物理遵循性、推理效率等核心指标上达到国际领先水平，进一步完善了国内 WAM 技术的全栈布局。

VLA 向 WAM 的跃迁本质是具身智能从” 模仿人类指令” 到” 理解物理因果” 的认知升级。世界模型的实时交互生成能力打破了仿真与现实的边界，为闭环规划与长程决策提供了统一框架。在下一阶段，WAM 将逐渐成为具身智能系统的核心组件。

5.2.2 数据范式的结构性变革

具身智能的数据生态正经历结构性变革：自我中心感知（Ego-centric Perception）、通用操作接口（UMI, Universal Manipulation Interface）、人类视频迁移学习（Human Video Transfer）、数据飞轮（Data Flywheel）与仿真到现实（Sim-to-Real）闭环的协同演化。

自我中心感知成为数据采集的主流形态。相较于传统的第三人称固定机位，自我中心视频天然携带以任务为中心的空间参考系，消除了视角变换带来的表征歧义。

2025 年下半年，Stanford 的 ARCap 与 DexCap 系统进一步将增强现实反馈与便携式动作捕捉相结合，实现了低成本、高保真的数据采集。预计 2026 年，Ego 数据将占据具身训练数据的 60% 以上。

通用操作接口（UMI）打破了数据采集的本体壁垒。UMI 通过手持式轻量化夹爪与便携式视觉追踪，实现了” 一次采集、跨本体复用” 的数据范式。其核心突破在于将数据采集从特定机器人平台解耦——同一组人类演示数据可通过策略迁移应用于不同构型的机械臂与夹爪，显著降低了数据收集的边际成本。2025 年，UMI 已展示出从桌面操作到移动双臂系统的无缝迁移能力。这种跨本体兼容性意味着具身数据将首次具备” 通用货币” 属性，打破” 一机一数据” 的孤岛状态，为构建大规模、跨平台的数据共享生态奠定基础。

人类视频迁移学习突破了机器人数据稀缺的根本约束。这一技术路径的成熟意味着：互联网规模的以自我为中心的人类视频将成为具身预训练的基础语料，显著降低对昂贵机器人数据采集的依赖。

数据飞轮机制实现了从” 数据驱动” 到” 数据自举” 的闭环。DexFlyWheel 展示了” 单条人类演示启动-残差强化学习微调-策略部署-数据增强” 的完整闭环，实现数据大规模扩展。2026 年，数据飞轮将成为具身系统部署的标准配置，推动模型能力持续增长。

大规模合成数据预训练验证了仿真到现实（Sim2Real）的新可能。上海人工智能实验室与北京大学联合发布的 InternData-A1 数据集首次证明，仅使用合成数据即可在 VLA 模型预训练中媲美使用真实数据集的最佳预训练模型的性能。该数据集包含超过 63 万条轨迹、7,433 小时数据，涵盖 4 种本体、18 项技能、70 个任务及 227 个场景，覆盖刚性、关节型、可变形及流体物体操作。通过高度自主、完全解耦的仿真流程，InternData-A1 实现了长程技能组合与跨本体数据生成，在 5 项真实世界任务和 4 项长程灵巧操作中展现出令人惊讶的零样本 Sim2Real 迁移能力。GigaBrain-0 等工作进一步引入世界模型生成数据，以高斯世界模型构建物理基础的合成环境，实现照片级真实感且物理可信的数据生成。2026 年，合成数据占比预计将继续提升，与真实数据形成互补。

多重数据变革共同指向” 低成本、可扩展、自增强” 的数据新范式。自我中心感知统一了人机数据采集形态，UMI 打破了本体数据孤岛，人类视频迁移突破了数据来源瓶颈，数据飞轮实现了能力增长的自我循环，而大规模合成数据预训练则打开了

数据规模化供给的终极通路。五者协同将具身智能的数据成本曲线从线性压向次线性，为大规模普及奠定数据基础。

5.2.3 技术范式演进与应用落地的发展

动作表示方法正经历从离散到连续、从独立到耦合的演进。流匹配（Flow Matching）已成为当前 VLA 模型的主流动作生成范式，VLA 模型的代表 π0.6 与世界模型的代表 Cosmos Policy 均采用流匹配的范式实现高频连续动作输出。

强化学习（RL）与 VLA 的深度整合是另一关键趋势。RLinf-VLA 框架首次实现流匹配 VLA 模型的在线强化学习微调，提出 Flow-Noise 与 Flow-SDE 两种微调方案，在 LIBERO 平台达到 98.3% 成功率。强化学习的引入，使 VLA 模型能够从真实环境反馈中持续学习，实现从模仿学习到自主学习的跨越。在后续研究中" 预训练 VLA+RL 后训练" 将成为具身领域主要研究方向之一。

长程任务的突破是具身智能落地的关键指标。随着具身智能算法架构与数据范式的持续成熟，领域技术迭代核心已从单一算法性能优化，转向构建全场景、全任务、全流程的自主作业能力。2026 年初，多项技术进展实现了任务全链路端到端架构闭环，完成了从实验室验证到实用化落地的关键跨越，在叠衣服这类长程任务的演示层出不穷。Pi0.6 模型证明了模型可以在新场景下通过强化学习收集快速收敛到一个新任务，进一步验证了具身智能系统在非结构化环境中的工程化可行性。

真正的落地还要考虑更多系统性问题，以家庭衣物处理场景为例，2026 年 1 月上线的元具身智能研究院发布的端到端大模型系统，实现了无人工干预下衣物识别、收拣、搬运、清洗前准备的全流程自主作业，突破了非结构化家庭环境中长时序任务闭环的核心技术瓶颈。该成果形成的技术方案可向餐具整理、物品归置等泛家庭轻家务场景迁移，为具身智能系统的多场景技术复用与能力拓展提供了标准化参考范式，推动个人服务机器人从科研验证向民用规模化应用加速演进。

5.3 具身智能研究平台发展展望

5.3.1 数据采集平台的便携化

具身数据采集硬件正经历从实验室专用设备向便携式、低成本、跨本体的转型。Stanford 提出的 UMI 数采系统是此方面的代表性工作，2025 年下半年，鹿明机器人

推出的 FastUMI Pro 背包式采集设备将单次采集成本降至 0.6 元以下，较传统遥操作降低一个数量级。

这些进展预示，2026 年具身数据采集将从” 专业设施” 转变为” 通用工具”，显著降低研究准入门槛。

5.3.2 仿真平台的开放化与标准化

具身仿真平台呈现百花齐放、加速融合的态势。NVIDIA Isaac Lab 与 Cosmos 的深度整合构建了从世界模型生成到策略训练的完整流水线；Genesis 物理引擎以高保真 GPU 加速渲染支持接触丰富交互；MuJoCo 与 Gymnasium 生态的持续扩展为算法验证提供标准化接口。国内方面，北京大学、清华大学、智元机器人等机构均推出了成熟的开源仿真平台，为具身智能算法的初步验证提供了低成本、高适配的研发环境。其中，智元机器人研发的 Genie Sim 开源仿真平台，为业内首个大语言模型驱动的开源具身仿真平台。平台融合三维重建与视觉生成技术构建高保真数字孪生物理环境，通过自研的大模型驱动场景泛化技术实现分钟级万级训练场景生成；同步开源全量代码、三维资产、仿真数据集与标准化评测工具，形成场景生成-策略训练-性能评测的全流程技术闭环，为具身智能算法的快速迭代与低成本验证提供标准化技术支撑。

实机验证平台的规模化建设成为弥合” 仿真-现实鸿沟” 的关键基础设施。全球首个大规模真机基准测试平台 RoboChallenge 支持 30 项复杂任务的远程接入验证；北京的国家地方共建人形机器人创新中心搭建的超 5000 平方米异构训练场，覆盖工业制造、民生服务等多场景，接入 100 余台不同构型机器人，实现数据采集、治理与模型验证的全流程闭环。

标准化场景设计与开放接口将成为实机平台的标配，使缺乏硬件资源的研发团队亦能开展真机测试。

5.3.3 数据生态的全球化与开源化

具身智能数据生态正经历从” 孤岛化” 向” 全球化” 的结构性转变。Open X-Embodiment 数据集持续扩展，AgiBot World 以 100 万条轨迹、2976 小时规模成为最大规模真实世界机器人操作数据集；Ego-Exo4D、Ego4D 等自我中心视频数据集为跨模态学习提供基础设施。开源社区方面，LeRobot 框架整合数据采集、训练与部署全流程，OpenVLA、

OpenPI 等开源模型使学术界与工业界得以在统一基线上开展研究。

数据标准的国际化协调日趋紧迫。随着中美欧在具身智能领域的竞争加剧，数据格式、评估协议、安全规范的标准化将成为全球协作与互操作的基础。开源化与全球化是具身智能跨越”数据壁垒”、实现能力跃迁的关键路径。统一的数据标准与开放的数据生态不仅降低重复建设成本，更为全球研究者的协同创新提供公共基础，其成熟度将直接影响 2026-2027 年具身智能技术的扩散速度与应用深度。

5.4 具身智能标准化发展展望

具身智能作为人工智能与物理实体深度融合的前沿领域，是培育未来产业、发展新质生产力的重要方向，标准化已成为引领技术创新、规范产业发展、保障安全应用的核心支撑。当前全球具身智能标准化整体处于起步探索阶段，尚未形成体系化布局，ISO/TC 299、IEC/TC59、IEC/TC61、IEC/TC62 等国际标准组聚焦机器人术语、性能测试、基础安全等方向布局标准，对具身智能特有的多模态感知、端到端决策、物理交互、智能体协同等核心属性覆盖不足，ISO/IEC JTC1 SC42 以人工智能为切入，围绕物理 AI 开展技术讨论，智能化标准与测评标准存在明显空白，我国在传统机器人领域国际标准话语权较弱，亟需通过具身智能标准化实现突破。

未来，具身智能标准化将进入快速完善、全面落地、深度赋能的新阶段，以标准引领技术迭代、产业协同与场景规模化应用，全面构建起全面覆盖人形机器人、仿生机器人、智能汽车等多元产品形态、面向制造、民生、电力、医疗、救援等重点行业的“具身智能 +”标准体系。标准供给上，将聚焦基础定义、智能化、接口适配、安全治理四大迫切方向，加快急需标准研制，形成国家标准、行业标准、团体标准协同配套的格局。基础通用方面，优先开展术语、测评方法、身份管理等基础标准，为行业提供统一标尺。智能化方面，推动端到端模型、运动控制、集群协同、类脑计算等核心技术与标准同步演进，规范“感知—决策—执行”全链路技术要求。在接口与适配方面，聚焦数据统一格式、算法与硬件的接口协议等标准，提升数据可复用性，破解“算法修改硬件、硬件适配算法”的重复开发困境，切实形成产业合力。在安全治理上，构建覆盖全生命周期的安全与伦理标准体系，明确机械安全、功能安全、数据安全、算法可信、伦理合规等要求，划定人机交互边界，防范技术风险与伦理争议，提升产业发展安全性与公众认可度。在国际合作方面，我国将持续深化与 ISO、IEC 等国际组织对接，推动优势标准向国际转化，积极参与国际规则制定，争取国际标准

化组织关键职务，在具身智能领域提交中国方案，提升全球标准话语权。

第六章　总结

具身智能作为人工智能领域的前沿研究方向，其本质特征体现在智能体能够通过与环境的动态交互，实现自主学习与能力演进。这一交叉学科融合了计算机科学、机器人学和认知神经科学等多领域知识体系，旨在构建具有环境感知、自主决策和精准执行能力的智能系统，使其能够在非结构化环境中依旧具备智能行为的能力。从理论框架到技术实现，具身智能的发展代表着当今人工智能研究范式正经历虚拟世界算法驱动到真实环境实体交互的重要转变，其核心在于建立“感知—认知—行动”的动态闭环机制，推动人工智能从“离身计算”向“具身智能”的范式跃迁。

具身智能目前形成了相对完备的技术体系。其中，多模态感知技术帮助实智能体实现对所处环境的可靠理解；基于大语言模型的认知推理系统则帮助智能体实现对复杂任务的语义理解和行为规划；而基于强化学习的方法则为智能体提供了基于环境反馈的策略自主优化。上述技术的协同创新有效增强了智能体在复杂开放环境中的适应能力，其中融合了感知、认知与执行的”视觉-语言-动作”模型正逐步成为具身智能的新一代核心技术框架。

高质量数据集是具身智能研究的基础。MuJoCo、NVIDIA Isaac 和 RoboVerse 等仿真平台通过持续优化物理引擎和计算效率，为智能体的训练和验证提供了可靠的实验环境。不过真实场景数据采集对于提升系统泛化性能和跨领域迁移能力仍具有不可替代的作用。近年来，随着遥操作系统等新型数据采集设备的应用，真实环境数据的获取效率得到显著提升。这些基础设施的不断完善，为具身智能技术的工程化应用提供了关键支撑。

具身智能目前正逐步渗透到各个行业。工业制造领域正在完成从传统刚性自动化向柔性智能化的转型升级；家用服务机器人正从单一功能模块向综合服务平台演进；医疗手术机器人通过精准控制显著提升了手术安全性和操作精度；农业智能装备则实现了基于环境感知的自主决策与精准作业。这些应用实践不仅带来了显著的生产效率提升，在作业安全、资源优化和可持续发展等方面也展现出重要的社会经济效益。

具身智能的未来发展主要聚焦于技术创新、应用拓展和产业生态三个维度：技术层面将致力于提升人机交互的自然性和智能体的环境适应能力；应用范围将向更

开放、更复杂的场景进行延伸；产业生态建设将聚焦标准规范体系的建立与完善。同时，具身智能的技术伦理和安全保障也将成为重要研究方向，以确保具身智能技术发展与社会需求的协调统一。

具身智能正处于从实验室研究向产业化应用过渡的关键阶段，机遇和挑战并存。一方面，具身核心算法的持续突破不断拓展着其应用边界；但另一方面，数据采集、算法泛化和系统可靠性仍然存在诸多瓶颈。因此需要产学研各方通力合作，围绕关键技术攻关、标准体系建设和产业生态培育，共同推进具身智能发展。可以预见，随着相关技术体系的持续完善和应用场景的不断丰富，具身智能将在未来十年内成为重塑人类生产生活方式的重要技术力量，为经济社会的发展提供新的增长动能。

参考文献

[1] Torne M, Simeonov A, Li Z, Chan A, Chen T, Gupta A, Agrawal P. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation [J]. arXiv preprint arXiv:2403.03949, 2024.

[2] Li X, Li J, Zhang Z, Zhang R, Jia F, Wang T, Fan H, Tseng K K, Wang R. Robogsim: A real2sim2real robotic gaussian splatting simulator [J]. arXiv preprint arXiv:2411.11839, 2024.

[3] Qureshi M N, Garg S, Yandun F, Held D, Kantor G, Silwal A. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting [J]. arXiv preprint arXiv:2409.10161, 2024.

[4] Han X, Liu M, Chen Y, Yu J, Lyu X, Tian Y, Wang B, Zhang W, Pang J. Re ˆ3 sim: Generating high-fidelity simulation data via 3d-photorealistic real-to-sim for robotic manipulation [J]. arXiv preprint arXiv:2502.08645, 2025.

[5] Lou H, Liu Y, Pan Y, Geng Y, Chen J, Ma W, Li C, Wang L, Feng H, Shi L, et al. Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation [J]. arXiv preprint arXiv:2408.14873, 2024.

[6] Wu Y, Pan L, Wu W, Wang G, Miao Y, Xu F, Wang H. Rl-gsbridge: 3d gaussian splatting based real2sim2real method for robotic manipulation learning [J]. arXiv preprint arXiv:2409.20291, 2024.

[7] Jiang Y, Wang C, Zhang R, Wu J, Fei-Fei L. Transic: Sim-to-real policy transfer by learning from online correction [J]. arXiv preprint arXiv:2405.10315, 2024.

[8] Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P. Domain randomization for transferring deep neural networks from simulation to the real world [C]//2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017: 23-30.

[9] Andrychowicz O M, Baker B, Chociej M, Jozefowicz R, McGrew B, Pachocki J, Petron A, Plappert M, Powell G, Ray A, et al. Learning dexterous in-hand manipulation [J]. The International Journal of Robotics Research, 2020, 39(1): 3-20.

[10] Matas J, James S, Davison A J. Sim-to-real reinforcement learning for deformable

object manipulation [C]//Conference on Robot Learning. PMLR, 2018: 734-743.

[11] Jiang Y, Yu C, Xie T, Li X, Feng Y, Wang H, Li M, Lau H, Gao F, Yang Y, et al. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality [C]//ACM SIGGRAPH 2024 Conference Papers. 2024: 1-1.

[12] Kaspar M, Osorio J D M, Bock J. Sim2real transfer for reinforcement learning without dynamics randomization [C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 4383-4388.

[13] Yu W, Tan J, Liu C K, Turk G. Preparing for the unknown: Learning a universal policy with online system identification [J]. arXiv preprint arXiv:1702.02453, 2017.

[14] Su K, Zhang X, Zhang S, Zhu J, Zhang B. To boost zero-shot generalization for embodied reasoning with vision-language pre-training [J]. IEEE Transactions on Image Processing, 2024.

[15] Yu A, Foote A, Mooney R, Martín-Martín R. Natural language can help bridge the sim2real gap [J]. arXiv preprint arXiv:2405.10020, 2024.

[16] He H, Huang J, Li Q, Wang X, Zhang F, Yang K, Meng L, Chu F. Maintagt: Sim2real-guided multimodal large model for intelligent maintenance with chain-of-thought reasoning [J]. arXiv preprint arXiv:2412.00481, 2024.

[17] Eftekhar A, Zeng K H, Duan J, Farhadi A, Kembhavi A, Krishna R. Selective visual representations improve convergence and generalization for embodied ai [J]. arXiv preprint arXiv:2311.04193, 2023.

[18] Sun H, Zhu F, Kong Y, Wang J, Zhao P. Continuous viewpoint planning in conjunction with dynamic exploration for active object recognition [J]. Entropy, 2021, 23(12): 1702.

[19] He L, Maiolino P. Embodied active tactile perception [C]//IOP Conference Series: Materials Science and Engineering: volume 1292. IOP Publishing, 2023: 012007.

[20] Chuang I, Lee A, Gao D, Naddaf-Sh M M, Soltani I. Active vision might be all you need: Exploring active vision in bimanual robotic manipulation [C]//2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025: 7952-7959.

[21] Qin Y, Zhou E, Liu Q, Yin Z, Sheng L, Zhang R, Qiao Y, Shao J. Mp5: A multi-modal open-ended embodied system in minecraft via active perception [C]//2024

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 16307-16316.

[22] Chen L, Zhan H, Chen K, Xu X, Yan Q, Cai C, Xu Y. Activegamer: Active gaussian mapping through efficient rendering [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025: 16486-16497.

[23] Schneider T, de Farias C, Calandra R, Chen L, Peters J. Apple: Toward general active perception via reinforcement learning [C]//International Conference on Learning Representations (ICLR). 2026.

[24] Wang Y, Du S, Xin Q, He Y, Qian W. Autonomous driving system driven by artificial intelligence perception fusion [J]. Academic Journal of Science and Technology, 2024, 9(2): 193-198.

[25] Xue T, Wang W, Ma J, Liu W, Pan Z, Han M. Progress and prospects of multimodal fusion methods in physical human–robot interaction: A review [J]. IEEE Sensors Journal, 2020, 20(18): 10355-10370.

[26] Xu D, Anguelov D, Jain A. Pointfusion: Deep sensor fusion for 3d bounding box estimation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 244-253.

[27] Pang S, Morris D, Radha H. Clocs: Camera-lidar object candidates fusion for 3d object detection [C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 10386-10393.

[28] Liu Z, Tang H, Amini A, Yang X, Mao H, Rus D L, Han S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation [C]//2023 IEEE international conference on robotics and automation (ICRA). IEEE, 2023: 2774-2781.

[29] Zhang W, Zhu S, Chen L, Bai L, Lam E Y, Guo E, Han J. High-resolution and real-time non-line-of-sight imaging based on spatial correlation [J]. Optics and Lasers in Engineering, 2025, 193: 109100.

[30] Braccini M. Metasensor: A proposal for sensor evolution in robotics [J]. Sensors, 2025, 25(3).

[31] Liu Q, Cui Y, Sun Z, Li G, Chen J, Ye Q. VTDexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning [C]//International Conference on Learning Representations (ICLR). 2025.

[32] Ye Q, Liu Q, Wang S, Chen J, Cui Y, Jin K, Chen H, Cai X, Li G, Chen J. Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity [J]. Science Robotics, 2026, 11(110): eady2869.

[33] Yu H, Cong Y, Sun G, Hou D, Liu Y, Dong J. Open-ended online learning for autonomous visual perception [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[34] Jing Y, Kong T. Learning to explore informative trajectories and samples for embodied perception [C]//2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 6050-6056.

[35] Li A, Ma X. Scalable cognitive developmental network: A strategy for integrating new perception online using relation evolution soinn [J]. Cognitive Systems Research, 2023, 79: 165-174.

[36] Liang X, Han A, Yan W, Raghunathan A, Abbeel P. Alp: Action-aware embodied learning for perception [J]. arXiv preprint arXiv:2306.10190, 2023.

[37] Chaplot D S, Dalal M, Gupta S, Malik J, Salakhutdinov R R. Seal: Self-supervised embodied active learning using exploration and 3d consistency [J]. Advances in neural information processing systems, 2021, 34: 13086-13098.

[38] Feng T, Wang X, Zhu W. Self-evolving embodied ai [J]. arXiv preprint arXiv:2602.04411, 2026.

[39] Tang Q, Zhang B, Liu J, Liu F, Liu Y. Dynamic token pruning in plain vision transformers for semantic segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 777-786.

[40] Cao J, Ye P, Li S, Yu C, Tang Y, Lu J, Chen T. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 15710-15719.

[41] Garavagno A M, Ragusa E, Frisoli A, Gastaldo P. An affordable hardware-aware neural architecture search for deploying convolutional neural networks on ultra-low-power computing platforms [J]. IEEE Sensors Letters, 2024, 8(5): 1-4.

[42] Garavagno A M, Ragusa E, Frisoli A, Gastaldo P. Searching neural architectures for sensor nodes on iot gateways [J]. IEEE Internet of Things Journal, 2025.

[43] Chehade A, Ragusa E, Gastaldo P, Zunino R. Hardware-aware neural architecture search for encrypted traffic classification on resource-constrained devices [J]. IEEE Transactions on Network and Service Management, 2026.

[44] Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, Finn C, Fu C, Gopalakrishnan K, Hausman K, et al. Do as i can, not as i say: Grounding language in robotic affordances [J]. arXiv preprint arXiv:2204.01691, 2022.

[45] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y. React: Synergizing reasoning and acting in language models [C]//International Conference on Learning Representations (ICLR). 2023.

[46] Lin K, Agia C, Migimatsu T, Pavone M, Bohg J. Text2motion: From natural language instructions to feasible plans [J]. Autonomous Robots, 2023, 47(8): 1345-1365.

[47] Liu R, Bai C, Lyu J, Sun S, Du Y, Li X. Vlp: Vision-language preference learning for embodied manipulation [J]. arXiv preprint arXiv:2502.11918, 2025.

[48] Liang J, Huang W, Xia F, Xu P, Hausman K, Ichter B, Florence P, Zeng A. Code as policies: Language model programs for embodied control [C]//2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 9493-9500.

[49] Mu Y, Chen J, Zhang Q, Chen S, Yu Q, Ge C, Chen R, Liang Z, Hu M, Tao C, et al. Robocodex: Multimodal code generation for robotic behavior synthesis [J]. arXiv preprint arXiv:2402.16117, 2024.

[50] Huang W, Wang C, Zhang R, Li Y, Wu J, Fei-Fei L. Voxposer: Composable 3d value maps for robotic manipulation with language models [J]. arXiv preprint arXiv:2307.05973, 2023.

[51] Pan M, Zhang J, Wu T, Zhao Y, Gao W, Dong H. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints [J]. arXiv preprint arXiv:2501.03841, 2025.

[52] Huang W, Wang C, Li Y, Zhang R, Fei-Fei L. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation [J]. arXiv preprint arXiv:2409.01652, 2024.

[53] Driess D, Xia F, Sajjadi M S, Lynch C, Chowdhery A, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, et al. Palm-e: An embodied multimodal language model [J]. 2023.

[54] Mu Y, Zhang Q, Hu M, Wang W, Ding M, Jin J, Wang B, Dai J, Qiao Y, Luo P. Em-

bodiedgpt: Vision-language pre-training via embodied chain of thought [J]. Advances in Neural Information Processing Systems, 2023, 36: 25081-25094.

[55] Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, Hamburger J, Jiang H, Liu M, Liu X, et al. Ego4d: Around the world in 3,000 hours of egocentric video [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 18995-19012.

[56] Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, Gopalakrishnan K, Hausman K, Herzog A, Hsu J, et al. Rt-1: Robotics transformer for real-world control at scale [J]. arXiv preprint arXiv:2212.06817, 2022.

[57] Zitkovich B, Yu T, Xu S, Xu P, Xiao T, Xia F, Wu J, Wohlhart P, Welker S, Wahid A, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control [C]//Conference on Robot Learning. PMLR, 2023: 2165-2183.

[58] O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, Lee A, Pooley A, Gupta A, Mandlekar A, Jain A, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0 [C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 6892-6903.

[59] Zhao X, Agrawal H, Batra D, Schwing A G. The surprising effectiveness of visual odometry techniques for embodied pointgoal navigation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16127-16136.

[60] Ramakrishnan S K, Chaplot D S, Al-Halah Z, Malik J, Grauman K. Poni: Potential functions for objectgoal navigation with interaction-free learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18890-18900.

[61] Anderson P, Wu Q, Teney D, Bruce J, Johnson M, Sünderhauf N, Reid I, Gould S, Van Den Hengel A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3674-3683.

[62] Krantz J, Gervet T, Yadav K, Wang A, Paxton C, Mottaghi R, Batra D, Malik J, Lee S, Chaplot D S. Navigating to objects specified by images [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 10916-10925.

[63] Fang K, Toshev A, Fei-Fei L, Savarese S. Scene memory transformer for embodied

agents in long-horizon tasks [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 538-547.

[64] Mousavian A, Toshev A, Fišer M, Košecká J, Wahid A, Davidson J. Visual representations for semantic target driven navigation [C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 8846-8852.

[65] Zhang S, Yu X, Song X, Wang X, Jiang S. Imagine before go: Self-supervised generative map for object goal navigation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16414-16425.

[66] Yokoyama N, Ha S, Batra D, Wang J, Bucher B. Vlfm: Vision-language frontier maps for zero-shot semantic navigation [C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 42-48.

[67] Li L H, Zhang P, Zhang H, Yang J, Li C, Zhong Y, Wang L, Yuan L, Zhang L, Hwang J N, et al. Grounded language-image pre-training [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10965-10975.

[68] Yin H, Xu X, Wu Z, Zhou J, Lu J. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation [J]. Advances in Neural Information Processing Systems, 2024, 37: 5285-5307.

[69] Sridhar A, Shah D, Glossop C, Levine S. Nomad: Goal masked diffusion policies for navigation and exploration [C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 63-70.

[70] Bar A, Zhou G, Tran D, Darrell T, LeCun Y. Navigation world models [J]. arXiv preprint arXiv:2412.03572, 2024.

[71] RoboCat: A self-improving robotic agent — deepmind.google [EB/OL]. https://deepmind.google/discover/blog/robocat-a-self-improving-robotic-agent/.

[72] Fu Z, Zhao Q, Wu Q, Wetzstein G, Finn C. Humanplus: Humanoid shadowing and imitation from humans [J]. arXiv preprint arXiv:2406.10454, 2024.

[73] Starting on the Right Foot with Reinforcement Learning | Boston Dynamics — bostondynamics.com [EB/OL]. https://bostondynamics.com/blog/starting-on-the-right-foot-with-reinforcement-learning/.

[74] 3 Questions: How the MIT mini cheetah learns to run — news.mit.edu [EB/OL].

https://news.mit.edu/2022/3-questions-how-mit-mini-cheetah-learns-run-fast-0317.

[75] Kumar A, Fu Z, Pathak D, Malik J. Rma: Rapid motor adaptation for legged robots [J]. arXiv preprint arXiv:2107.04034, 2021.

[76] Escontrela A, Peng X B, Yu W, Zhang T, Iscen A, Goldberg K, Abbeel P. Adversarial motion priors make good substitutes for complex reward functions [C]//2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022: 25-32.

[77] Xue Y, Dong W, Liu M, Zhang W, Pang J. A unified and general humanoid whole-body controller for fine-grained locomotion [J]. arXiv preprint arXiv:2502.03206, 2025.

[78] Multi-agent reinforcement learning for resource allocation in large-scale robotic warehouse sortation centers — amazon.science [EB/OL]. https://www.amazon.science/publications/multi-agent-reinforcement-learning-for-resource-allocation-in-large-scale-robotic-warehouse-sortati

[79] Chen J, Li X, Cao J, Zhu Z, Dong W, Liu M, Wen Y, Yu Y, Zhang L, Zhang W. Rhino: Learning real-time humanoid-human-object interaction from human demonstrations [J]. arXiv preprint arXiv:2502.13134, 2025.

[80] Dourish P. Where the action is: the foundations of embodied interaction [M]. MIT press, 2001.

[81] Gao X, Gao Q, Gong R, Lin K, Thattai G, Sukhatme G S. Dialfred: Dialogue-enabled agents for embodied instruction following [J]. IEEE Robotics and Automation Letters, 2022, 7(4): 10049-10056.

[82] Kalinowska A, Pilarski P M, Murphey T D. Embodied communication: How robots and people communicate through physical interaction [J]. Annual review of control, robotics, and autonomous systems, 2023, 6(1): 205-232.

[83] Padmakumar A, Thomason J, Shrivastava A, Lange P, Narayan-Chen A, Gella S, Piramuthu R, Tur G, Hakkani-Tur D. Teach: Task-driven embodied agents that chat [C]// Proceedings of the AAAI Conference on Artificial Intelligence: volume 36. 2022: 2017-2025.

[84] Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choromanski K, Ding T, Driess D, Dubey A, Finn C, et al. Rt-2: Vision-language-action models transfer web

knowledge to robotic control [J]. arXiv preprint arXiv:2307.15818, 2023.

[85] Retzlaff C O, Das S, Wayllace C, Mousavi P, Afshari M, Yang T, Saranti A, Angerschmid A, Taylor M E, Holzinger A. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities [J]. Journal of Artificial Intelligence Research, 2024, 79: 359-415.

[86] Long Y, Wei W, Huang T, Wang Y, Dou Q. Human-in-the-loop embodied intelligence with interactive simulation environment for surgical robot learning [J]. IEEE Robotics and Automation Letters, 2023, 8(8): 4441-4448.

[87] Li Z, Wu W, Guo Y, Sun J, Han Q L. Embodied multi-agent systems: A review [J/OL]. IEEE/CAA Journal of Automatica Sinica, 2025, 12(6): 1095-1116. DOI: 10.1109/JAS.2025.125552.

[88] Zhou X, Wen X, Wang Z, Gao Y, Li H, Wang Q, Yang T, Lu H, Cao Y, Xu C, Gao F. Swarm of micro flying robots in the wild [J/OL]. Science Robotics, 2022, 7(66): eabm5954. DOI: 10.1126/scirobotics.abm5954.

[89] Tu Y, Lam T L. Configuration identification for a freeform modular self-reconfigurable robot - freesn [J/OL]. IEEE Transactions on Robotics, 2023, 39(6): 4636-4652. DOI: 10.1109/TRO.2023.3303848.

[90] Tan H, Hao X, Chi C, Lin M, Lyu Y, Cao M, Liang D, Chen Z, Lyu M, Peng C, et al. Roboos: A hierarchical embodied framework for cross-embodiment and multi-agent collaboration [J]. arXiv preprint arXiv:2505.03673, 2025.

[91] Zhang X, Qin H, Wang F, Dong Y, Li J. Lamma-p: Generalizable multi-agent long-horizon task allocation and planning with lm-driven pddl planner [C]//2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025: 10221-10221.

[92] Gao M, Li J, Lin Y, Liu X, Ji J, Pan X, Xu Z, Li X, Li M, Ji W, et al. Arcadia: Toward a full-lifecycle framework for embodied lifelong learning [J]. arXiv preprint arXiv:2512.00076, 2025.

[93] Wang X, Dong J, Liu B, Lyu Q, Liu L, Han Z. Lifelong embodied navigation learning [J]. arXiv preprint arXiv:2603.06073, 2026.

[94] Chen X, Gao Y, Liu H, Yang F, Ghadirzadeh A, Yang J, Liang B, Zhang C, Lam T L, Zhu S C. Cross-robot behavior adaptation through intention alignment [J/OL]. Sci-

ence Robotics, 2026, 11(112): eadv2250. https://www.science.org/doi/abs/10.1126/scirobotics.adv2250.

[95] Brown T, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[96] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A. Emerging properties in self-supervised vision transformers [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9650-9660.

[97] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg A C, Lo W Y, et al. Segment anything [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 4015-4026.

[98] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, et al. Dinov2: Learning robust visual features without supervision [J]. arXiv preprint arXiv:2304.07193, 2023.

[99] Ravi N, Gabeur V, Hu Y T, Hu R, Ryali C, Ma T, Khedr H, Rädle R, Rolland C, Gustafson L, et al. Sam 2: Segment anything in images and videos [J]. arXiv preprint arXiv:2408.00714, 2024.

[100] LeCun Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27 [J]. Open Review, 2022, 62(1): 1-62.

[101] Chen B, Xia F, Ichter B, Rao K, Gopalakrishnan K, Ryoo M S, Stone A, Kappler D. Open-vocabulary queryable scene representations for real world planning [C]//2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 11509-11522.

[102] Tang Y, Yu W, Tan J. Heiga zen, aleksandra faust, and tatsuya harada [J]. SayTap: Language to quadrupedal locomotion, 2023, 2.

[103] Bommasani R, Hudson D A, Adeli E, Altman R, Arora S, von Arx S, Bernstein M S, Bohg J, Bosselut A, Brunskill E, et al. On the opportunities and risks of foundation models [J]. arXiv preprint arXiv:2108.07258, 2021.

[104] Ha D, Schmidhuber J. World models [J]. arXiv preprint arXiv:1803.10122, 2018.

[105] Finn C, Goodfellow I, Levine S. Unsupervised learning for physical interaction through video prediction [J]. Advances in neural information processing systems,

2016, 29.

[106] Finn C, Levine S. Deep visual foresight for planning robot motion [C]//2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017: 2786-2793.

[107] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[108] Wang A, Chen H, Lin Z, Han J, Ding G. Repvit: Revisiting mobile cnn from vit perspective [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 15909-15920.

[109] Shi H, Xu H, Huang Z, Li Y, Wu J. Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks [J]. The International Journal of Robotics Research, 2024, 43(4): 533-549.

[110] Qi C R, Su H, Mo K, Guibas L J. Pointnet: Deep learning on point sets for 3d classification and segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[111] Qian G, Li Y, Peng H, Mai J, Hammoud H, Elhoseiny M, Ghanem B. Pointnext: Revisiting pointnet++ with improved training and scaling strategies [J]. Advances in neural information processing systems, 2022, 35: 23192-23204.

[112] Gornet J, Thomson M. Automated construction of cognitive maps with visual predictive coding [J]. Nature Machine Intelligence, 2024, 6(7): 820-833.

[113] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al. Llama: Open and efficient foundation language models [J]. arXiv preprint arXiv:2302.13971, 2023.

[114] Du N, Huang Y, Dai A M, Tong S, Lepikhin D, Xu Y, Krikun M, Zhou Y, Yu A W, Firat O, et al. Glam: Efficient scaling of language models with mixture-of-experts [C]//International conference on machine learning. PMLR, 2022: 5547-5569.

[115] Mu F, Shi L, Wang S, Yu Z, Zhang B, Wang C, Liu S, Wang Q. Clarifygpt: Empowering llm-based code generation with intention clarification [J]. arXiv preprint arXiv:2310.10996, 2023.

[116] Gestrin E, Kuhlmann M, Seipp J. Nl2plan: Robust llm-driven planning from minimal

text descriptions [J]. arXiv preprint arXiv:2405.04215, 2024.

[117] Hua P, Liu M, Macaluso A, Lin Y, Zhang W, Xu H, Wang L. Gensim2: Scaling robot data generation with multi-modal and reasoning llms [J]. arXiv preprint arXiv:2410.03645, 2024.

[118] Wang L, Ling Y, Yuan Z, Shridhar M, Bao C, Qin Y, Wang B, Xu H, Wang X. Gensim: Generating robotic simulation tasks via large language models [J]. arXiv preprint arXiv:2310.01361, 2023.

[119] Jang E, Irpan A, Khansari M, Kappler D, Ebert F, Lynch C, Levine S, Finn C. Bc-z: Zero-shot task generalization with robotic imitation learning [C]//Conference on Robot Learning. PMLR, 2022: 991-1002.

[120] Lin K, Agia C, Migimatsu T, Pavone M, Bohg J. Text2motion: From natural language instructions to feasible plans [J]. Autonomous Robots, 2023, 47(8): 1345-1365.

[121] Xie Q, Zhang T, Xu K, Johnson-Roberson M, Bisk Y. Reasoning about the unseen for efficient outdoor object navigation [J]. arXiv preprint arXiv:2309.10103, 2023.

[122] Chen J, Li G, Kumar S, Ghanem B, Yu F. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers [J]. arXiv preprint arXiv:2305.16925, 2023.

[123] Xu J, Tian Y, Ma P, Rus D, Sueda S, Matusik W. Prediction-guided multi-objective reinforcement learning for continuous robot control [C]//International conference on machine learning. PMLR, 2020: 10607-10616.

[124] Salzmann T, Ivanovic B, Chakravarty P, Pavone M. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data [C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer, 2020: 683-700.

[125] Esser P, Chiu J, Atighehchian P, Granskog J, Germanidis A. Structure and content-guided video synthesis with diffusion models [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 7346-7356.

[126] Yu L, Cheng Y, Sohn K, Lezama J, Zhang H, Chang H, Hauptmann A G, Yang M H, Hao Y, Essa I, et al. Magvit: Masked generative video transformer [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10459-10469.

[127] Escontrela A, Adeniji A, Yan W, Jain A, Peng X B, Goldberg K, Lee Y, Hafner D, Abbeel P. Video prediction models as rewards for reinforcement learning [J]. Advances in Neural Information Processing Systems, 2023, 36: 68760-68783.

[128] Bruce J, Dennis M D, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steigerwald R, Apps C, et al. Genie: Generative interactive environments [C]//Forty-first International Conference on Machine Learning. 2024.

[129] Wu H, Jing Y, Cheang C, Chen G, Xu J, Li X, Liu M, Li H, Kong T. Unleashing large-scale video generative pre-training for visual robot manipulation [J]. arXiv preprint arXiv:2312.13139, 2023.

[130] Cheang C L, Chen G, Jing Y, Kong T, Li H, Li Y, Liu Y, Wu H, Xu J, Yang Y, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation [J]. arXiv preprint arXiv:2410.06158, 2024.

[131] Du Y, Yang S, Dai B, Dai H, Nachum O, Tenenbaum J, Schuurmans D, Abbeel P. Learning universal policies via text-guided video generation [J]. Advances in neural information processing systems, 2023, 36: 9156-9172.

[132] Zhou S, Du Y, Chen J, Li Y, Yeung D Y, Gan C. Robodreamer: Learning compositional world models for robot imagination [J]. arXiv preprint arXiv:2404.12377, 2024.

[133] He H, Bai C, Pan L, Zhang W, Zhao B, Li X. Large-scale actionless video pre-training via discrete diffusion for efficient policy learning [J]. arXiv e-prints, 2024: arXiv-2402.

[134] Feng Y, Han J, Yang Z, Yue X, Levine S, Luo J. Reflective planning: Vision-language models for multi-stage long-horizon robotic manipulation [J]. arXiv preprint arXiv:2502.16707, 2025.

[135] Lin J, Liu L, Lu D, Jia K. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 27906-27916.

[136] Qian S, Chen W, Bai M, Zhou X, Tu Z, Li L E. Affordancellm: Grounding affordance from vision language models [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 7587-7597.

[137] Ye Y, Li X, Gupta A, De Mello S, Birchfield S, Song J, Tulsiani S, Liu S. Affordance diffusion: Synthesizing hand-object interactions [C]//Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. 2023: 22479-22489.

[138] Huang W, Wang C, Li Y, Zhang R, Fei-Fei L. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation [J]. arXiv preprint arXiv:2409.01652, 2024.

[139] Yuan Z, Xue Z, Yuan B, Wang X, Wu Y, Gao Y, Xu H. Pre-trained image encoder for generalizable visual reinforcement learning [J]. Advances in Neural Information Processing Systems, 2022, 35: 13022-13037.

[140] Parisi S, Rajeswaran A, Purushwalkam S, Gupta A. The unsurprising effectiveness of pre-trained vision models for control [C]//international conference on machine learning. PMLR, 2022: 17359-17371.

[141] Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, Hamburger J, Jiang H, Liu M, Liu X, et al. Ego4d: Around the world in 3,000 hours of egocentric video [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 18995-19012.

[142] Goyal R, Ebrahimi Kahou S, Michalski V, Materzynska J, Westphal S, Kim H, Haenel V, Fruend I, Yianilos P, Mueller-Freitag M, et al. The" something something" video database for learning and evaluating visual common sense [C]//Proceedings of the IEEE international conference on computer vision. 2017: 5842-5850.

[143] Nair S, Rajeswaran A, Kumar V, Finn C, Gupta A. R3m: A universal visual representation for robot manipulation [J]. arXiv preprint arXiv:2203.12601, 2022.

[144] Karamcheti S, Nair S, Chen A S, Kollar T, Finn C, Sadigh D, Liang P. Language-driven representation learning for robotics [J]. arXiv preprint arXiv:2302.12766, 2023.

[145] Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, Finn C, Fu C, Gopalakrishnan K, Hausman K, et al. Do as i can, not as i say: Grounding language in robotic affordances [J]. arXiv preprint arXiv:2204.01691, 2022.

[146] Huang W, Xia F, Xiao T, Chan H, Liang J, Florence P, Zeng A, Tompson J, Mordatch I, Chebotar Y, et al. Inner monologue: Embodied reasoning through planning with language models [J]. arXiv preprint arXiv:2207.05608, 2022.

[147] Xie S M, Pham H, Dong X, Du N, Liu H, Lu Y, Liang P S, Le Q V, Ma T, Yu A W. Doremi: Optimizing data mixtures speeds up language model pretraining [J]. Advances in Neural Information Processing Systems, 2023, 36: 69798-69818.

[148] Song C H, Wu J, Washington C, Sadler B M, Chao W L, Su Y. Llm-planner: Few-shot grounded planning for embodied agents with large language models [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 2998-3009.

[149] Chen Y, Cui W, Chen Y, Tan M, Zhang X, Zhao D, Wang H. Robogpt: an intelligent agent of making embodied long-term decisions for daily instruction tasks [J]. arXiv preprint arXiv:2311.15649, 2023.

[150] Rana K, Haviland J, Garg S, Abou-Chakra J, Reid I, Suenderhauf N. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning [J]. arXiv preprint arXiv:2307.06135, 2023.

[151] Hu Y, Lin F, Zhang T, Yi L, Gao Y. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning [J]. arXiv preprint arXiv:2311.17842, 2023.

[152] Skreta M, Zhou Z, Yuan J L, Darvish K, Aspuru-Guzik A, Garg A. Replan: Robotic replanning with perception and language models [J]. arXiv preprint arXiv:2401.04157, 2024.

[153] Zhao T Z, Kumar V, Levine S, Finn C. Learning fine-grained bimanual manipulation with low-cost hardware [J]. arXiv preprint arXiv:2304.13705, 2023.

[154] Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, Gopalakrishnan K, Hausman K, Herzog A, Hsu J, et al. Rt-1: Robotics transformer for real-world control at scale [J]. arXiv preprint arXiv:2212.06817, 2022.

[155] Vuong Q, Levine S, Walke H R, Pertsch K, Singh A, Doshi R, Xu C, Luo J, Tan L, Shah D, et al. Open x-embodiment: Robotic learning datasets and rt-x models [C]// Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023. 2023.

[156] Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, Rafailov R, Foster E, Lam G, Sanketi P, et al. Openvla: An open-source vision-language-action model [J]. arXiv preprint arXiv:2406.09246, 2024.

[157] Li X, Liu M, Zhang H, Yu C, Xu J, Wu H, Cheang C, Jing Y, Zhang W, Liu H, et al. Vision-language foundation models as effective robot imitators [J]. arXiv preprint arXiv:2311.01378, 2023.

[158] Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Mil-

lican K, Reynolds M, et al. Flamingo: a visual language model for few-shot learning [J]. Advances in neural information processing systems, 2022, 35: 23716-23736.

[159] Li X, Liu M, Zhang H, Yu C, Xu J, Wu H, Cheang C, Jing Y, Zhang W, Liu H, et al. Vision-language foundation models as effective robot imitators [J]. arXiv preprint arXiv:2311.01378, 2023.

[160] Team O M, Ghosh D, Walke H, Pertsch K, Black K, Mees O, Dasari S, Hejna J, Kreiman T, Xu C, et al. Octo: An open-source generalist robot policy [J]. arXiv preprint arXiv:2405.12213, 2024.

[161] Qu D, Song H, Chen Q, Yao Y, Ye X, Ding Y, Wang Z, Gu J, Zhao B, Wang D, et al. Spatialvla: Exploring spatial representations for visual-language-action model [J]. arXiv preprint arXiv:2501.15830, 2025.

[162] Shi L X, Ichter B, Equi M, Ke L, Pertsch K, Vuong Q, Tanner J, Walling A, Wang H, Fusai N, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models [J]. arXiv preprint arXiv:2502.19417, 2025.

[163] Chi C, Xu Z, Feng S, Cousineau E, Du Y, Burchfiel B, Tedrake R, Song S. Diffusion policy: Visuomotor policy learning via action diffusion [J]. The International Journal of Robotics Research, 2023: 02783649241273668.

[164] Ze Y, Zhang G, Zhang K, Hu C, Wang M, Xu H. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations [J]. arXiv preprint arXiv:2403.03954, 2024.

[165] Black K, Brown N, Driess D, Esmail A, Equi M, Finn C, Fusai N, Groom L, Hausman K, Ichter B, et al. $\pi 0$: A vision-language-action flow model for general robot control, 2024 [J]. URL https://arxiv. org/abs/2410.24164.

[166] Carlini N, Mishra P, Vaidya T, Zhang Y, Sherr M, Shields C, Wagner D, Zhou W. Hidden voice commands [C]//25th USENIX security symposium (USENIX security 16). 2016: 513-530.

[167] Yan C, Zhang G, Ji X, Zhang T, Zhang T, Xu W. The feasibility of injecting inaudible voice commands to voice assistants [J]. IEEE Transactions on Dependable and Secure Computing, 2019, 18(3): 1108-1124.

[168] Schönherr L, Kohls K, Zeiler S, Holz T, Kolossa D. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding [J]. arXiv preprint

arXiv:1808.05665, 2018.

[169] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text [C]//2018 IEEE security and privacy workshops (SPW). IEEE, 2018: 1-7.

[170] Zhang H, Zhu C, Wang X, Zhou Z, Yin C, Li M, Xue L, Wang Y, Hu S, Liu A, et al. Badrobot: Jailbreaking embodied llms in the physical world [C]//The Thirteenth International Conference on Learning Representations. 2024.

[171] Robey A, Ravichandran Z, Kumar V, Hassani H, Pappas G J. Jailbreaking llm-controlled robots [J]. arXiv preprint arXiv:2410.13691, 2024.

[172] Lu X, Huang Z, Li X, Xu W, et al. Poex: Policy executable embodied ai jailbreak attacks [J]. arXiv preprint arXiv:2412.16633, 2024.

[173] Liu A, Zhou Y, Liu X, Zhang T, Liang S, Wang J, Pu Y, Li T, Zhang J, Zhou W, et al. Compromising embodied agents with contextual backdoor attacks [J]. arXiv preprint arXiv:2408.02882, 2024.

[174] Jiao R, Xie S, Yue J, Sato T, Wang L, Wang Y, Chen Q A, Zhu Q. Can we trust embodied agents? exploring backdoor attacks against embodied llm-based decision-making systems [J]. arXiv preprint arXiv:2405.20774, 2024.

[175] Ren A Z, Dixit A, Bodrova A, Singh S, Tu S, Brown N, Xu P, Takayama L, Xia F, Varley J, et al. Robots that ask for help: Uncertainty alignment for large language model planners [J]. arXiv preprint arXiv:2307.01928, 2023.

[176] Liang K, Zhang Z, Fisac J F. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity [J]. Advances in Neural Information Processing Systems, 2024, 37: 71998-72031.

[177] Park J, Lim S, Lee J, Park S, Chang M, Yu Y, Choi S. Clara: classifying and disambiguating user commands for reliable interactive robotic agents [J]. IEEE Robotics and Automation Letters, 2023, 9(2): 1059-1066.

[178] Chen L, Wang L, Dong H, Du Y, Yan J, Yang F, Li S, Zhao P, Qin S, Rajmohan S, et al. Introspective tips: Large language model for in-context decision making [J]. arXiv preprint arXiv:2305.11598, 2023.

[179] Sathyamoorthy D, Fitry Z, Selamat E, Hassan S, Firdaus A, Zaimy Z. Evaluation of the vulnerabilities of unmanned aerial vehicles (uavs) to global positioning system (gps) jamming and spoofing [J]. Defence S and T Technical Bulletin, 2020, 13: 333-343.

[180] Elezi E, Çankaya G, Boyacı A, Yarkan S. The effect of electronic jammers on gps signals [C]//2019 16th International Multi-Conference on Systems, Signals & Devices (SSD). IEEE, 2019: 652-656.

[181] He D, Liu H, Chan S, Guizani M. How to govern the non-cooperative amateur drones? [J]. IEEE Network, 2019, 33(3): 184-189.

[182] Shen J, Won J Y, Chen Z, Chen Q A. Drift with devil: Security of {Multi-Sensor} fusion based localization in {High-Level} autonomous driving under {GPS} spoofing [C]//29th USENIX security symposium (USENIX Security 20). 2020: 931-948.

[183] Son Y, Shin H, Kim D, Park Y, Noh J, Choi K, Choi J, Kim Y. Rocking drones with intentional sound noise on gyroscopic sensors [C]//24th USENIX security symposium (USENIX Security 15). 2015: 881-896.

[184] Trippel T, Weisse O, Xu W, Honeyman P, Fu K. Walnut: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks [C]//2017 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2017: 3-18.

[185] Liu W, Ren G, Yu R, Guo S, Zhu J, Zhang L. Image-adaptive yolo for object detection in adverse weather conditions [C]//Proceedings of the AAAI conference on artificial intelligence: volume 36. 2022: 1792-1800.

[186] Qian C, Guo Y, Mo Y, Li W. Weatherdg: Llm-assisted procedural weather generation for domain-generalized semantic segmentation [J]. IEEE Robotics and Automation Letters, 2025.

[187] Xu X, Zhang J, Li Y, Wang Y, Yang Y, Shen H T. Adversarial attack against urban scene segmentation for autonomous vehicles [J]. IEEE Transactions on Industrial Informatics, 2020, 17(6): 4117-4126.

[188] Chen M, Tu J, Qi C, Dang Y, Zhou F, Wei W, Yin J. Towards physically-realizable adversarial attacks in embodied vision navigation [J]. arXiv preprint arXiv:2409.10071, 2024.

[189] 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.

[190] Nassi B, Mirsky Y, Nassi D, Ben-Netanel R, Drokin O, Elovici Y. Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks [C]//

Proceedings of the 2020 ACM SIGSAC conference on computer and communications security. 2020: 293-308.

[191] Tu J, Ren M, Manivasagam S, Liang M, Yang B, Du R, Cheng F, Urtasun R. Physically realizable adversarial examples for lidar object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 13716-13725.

[192] Zhu Y, Miao C, Zheng T, Hajiaghajani F, Su L, Qiao C. Can we use arbitrary objects to attack lidar perception in autonomous driving? [C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021: 1945-1960.

[193] Jin Z, Ji X, Cheng Y, Yang B, Yan C, Xu W. Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle [C]//2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023: 1822-1839.

[194] Li Y, Wen C, Juefei-Xu F, Feng C. Fooling lidar perception via adversarial trajectory perturbation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7898-7907.

[195] Yan C, Xu W, Liu J. Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle [J]. Def Con, 2016, 24(8): 109.

[196] Lim B S, Keoh S L, Thing V L. Autonomous vehicle ultrasonic sensor vulnerability and impact assessment [C]//2018 IEEE 4th World Forum on Internet of Things (WF-IoT). IEEE, 2018: 231-236.

[197] Srinivasan K, Eysenbach B, Ha S, Tan J, Finn C. Learning to be safe: Deep rl with a safety critic [J]. arXiv preprint arXiv:2010.14603, 2020.

[198] Allamaa J P, Patrinos P, Ohtsuka T, Son T D. Real-time mpc with control barrier functions for autonomous driving using safety enhanced collocation [J]. IFAC-PapersOnLine, 2024, 58(18): 392-399.

[199] Xiao W, Wang T H, Gan C, Rus D. Safediffuser: Safe planning with diffusion probabilistic models [J]. arXiv preprint arXiv:2306.00148, 2023.

[200] Lyu Y, Luo W, Dolan J M. Probabilistic safety-assured adaptive merging control for autonomous vehicles [C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Ieee, 2021: 10764-10770.

[201] Zhao W, He T, Liu C. Model-free safe control for zero-violation reinforcement learn-

ing [C]//5th Annual Conference on Robot Learning. 2021.

[202] Xie Y, Guo X, Wang C, Liu K, Chen L. Advdiffuser: Generating adversarial safety-critical driving scenarios via guided diffusion [C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024: 9983-9989.

[203] Abeysirigoonawardena Y, Shkurti F, Dudek G. Generating adversarial driving scenarios in high-fidelity simulators [C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 8271-8277.

[204] Xu S, Mi L, Gilpin L H. A framework for generating dangerous scenes for testing robustness [C]//Progress and Challenges in Building Trustworthy Embodied AI. 2022.

[205] Jia Y, Poskitt C M, Sun J, Chattopadhyay S. Physical adversarial attack on a robotic arm [J]. IEEE Robotics and Automation Letters, 2022, 7(4): 9334-9341.

[206] Li J, Zhu Y, Xu Z, Gu J, Zhu M, Liu X, Liu N, Peng Y, Feng F, Tang J. Mmro: Are multimodal llms eligible as the brain for in-home robotics? [J]. arXiv preprint arXiv:2406.19693, 2024.

[207] Kirschner R J, Kurdas A, Karacan K, Junge P, Birjandi S A B, Mansfeld N, Abdolshah S, Haddadin S. Towards a reference framework for tactile robot performance and safety benchmarking [C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021: 4290-4297.

[208] Yu X, Wu J, Xu C, Luo H, Ou L. Adaptive human-robot collaboration control based on optimal admittance parameters [J]. Journal of Shanghai Jiaotong University (Science), 2022, 27(5): 589-601.

[209] Wang X, Pan H, Zhang H, Li M, Hu S, Zhou Z, Xue L, Guo P, Wang Y, Wan W, et al. Trojanrobot: Physical-world backdoor attacks against vlm-based robotic manipulation [J]. arXiv preprint arXiv:2411.11683, 2024.

[210] Hancock A J, Ren A Z, Majumdar A. Run-time observation interventions make vision-language-action models more visually robust [J]. arXiv preprint arXiv:2410.01971, 2024.

[211] Katayama S, Takasugi N, Kaneko M, Nagatsuka N, Kinoshita M. Robustifying model-based locomotion by zero-order stochastic nonlinear model predictive control with guard saltation matrix [C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024: 11932-11939.

[212] Siekmann J, Green K, Warila J, Fern A, Hurst J. Blind bipedal stair traversal via sim-to-real reinforcement learning [J]. arXiv preprint arXiv:2105.08328, 2021.

[213] Zhu S, Huang R, Mou L, Zhao H. Robust robot walker: Learning agile locomotion over tiny traps [J]. arXiv preprint arXiv:2409.07409, 2024.

[214] Zhang Q, Jin S, Zhu R, Sun J, Zhang X, Chen Q A, Mao Z M. On data fabrication in collaborative vehicular perception: Attacks and countermeasures [C]//33rd USENIX Security Symposium (USENIX Security 24). 2024: 6309-6326.

[215] Tu J, Wang T, Wang J, Manivasagam S, Ren M, Urtasun R. Adversarial attacks on multi-agent communication [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7768-7777.

[216] Chen Y, Zheng Z, Gong X. Marnet: Backdoor attacks against cooperative multi-agent reinforcement learning [J]. IEEE Transactions on Dependable and Secure Computing, 2022, 20(5): 4188-4198.

[217] Zheng X, Ma X, Wang S, Wang X, Shen C, Wang C. Toward evaluating robustness of reinforcement learning with adversarial policy [C]//2024 54th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, 2024: 288-301.

[218] Wang S, Cheng X, Lam T L, Zhang T. Mobile cooperative robot safe interaction method based on embodied perception [C]//2024 IEEE 18th International Conference on Control & Automation (ICCA). IEEE, 2024: 692-698.

[219] Schepp S R, Thumm J, Liu S B, Althoff M. Sara: A tool for safe human-robot coexistence and collaboration through reachability analysis [C]//2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 4312-4317.

[220] Chen L, Chen L, Chen X, Lu H, Zheng Y, Wu J, Wang Y, Zhang Z, Xiong R. Compliance while resisting: A shear-thickening fluid controller for physical human-robot interaction [J]. The International Journal of Robotics Research, 2024, 43(11): 1731-1769.

[221] Jiang S, Wong L L. A hierarchical framework for robot safety using whole-body tactile sensors [C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 8021-8028.

[222] Zhao T Z, Kumar V, Levine S, Finn C. Learning fine-grained bimanual manipulation

with low-cost hardware [J]. arXiv preprint arXiv:2304.13705, 2023.

[223] Chi C, Xu Z, Pan C, Cousineau E, Burchfiel B, Feng S, Tedrake R, Song S. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots [J]. arXiv preprint arXiv:2402.10329, 2024.

[224] Mandlekar A, Zhu Y, Garg A, Booher J, Spero M, Tung A, Gao J, Emmons J, Gupta A, Orbay E, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation [C]//Conference on Robot Learning. PMLR, 2018: 879-893.

[225] Dasari S, Ebert F, Tian S, Nair S, Bucher B, Schmeckpeper K, Singh S, Levine S, Finn C. Robonet: Large-scale multi-robot learning [J]. arXiv preprint arXiv:1910.11215, 2019.

[226] Wu K, Hou C, Liu J, Che Z, Ju X, Yang Z, Li M, Zhao Y, Xu Z, Yang G, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation [J]. arXiv preprint arXiv:2412.13877, 2024.

[227] Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, Gopalakrishnan K, Hausman K, Herzog A, Hsu J, et al. Rt-1: Robotics transformer for real-world control at scale [J]. arXiv preprint arXiv:2212.06817, 2022.

[228] Fang H S, Fang H, Tang Z, Liu J, Wang C, Wang J, Zhu H, Lu C. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot [J]. arXiv preprint arXiv:2307.00595, 2023.

[229] Walke H R, Black K, Zhao T Z, Vuong Q, Zheng C, Hansen-Estruch P, He A W, Myers V, Kim M J, Du M, et al. Bridgedata v2: A dataset for robot learning at scale [C]//Conference on Robot Learning. PMLR, 2023: 1723-1736.

[230] Bharadhwaj H, Vakil J, Sharma M, Gupta A, Tulsiani S, Kumar V. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking [C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 4788-4795.

[231] Khazatsky A, Pertsch K, Nair S, Balakrishna A, Dasari S, Karamcheti S, Nasiriany S, Srirama M K, Chen L Y, Ellis K, et al. Droid: A large-scale in-the-wild robot manipulation dataset [J]. arXiv preprint arXiv:2403.12945, 2024.

[232] O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, Lee A, Pooley A, Gupta A, Mandlekar A, Jain A, et al. Open x-embodiment: Robotic learning datasets and rt-x

models: Open x-embodiment collaboration 0 [C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 6892-6903.

[233] OpenDriveLab. Github - opendrivelab/agibot-world: The large-scale manipulation platform for scalable and intelligent embodied systems [EB/OL]. 2025. https://github.com/OpenDriveLab/AgiBot-World.

[234] Hou C, Wu K, Liu J, Che Z, Wu D, Liao F, Li G, He J, Feng Q, Jin Z, et al. Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence [J]. arXiv preprint arXiv:2512.24653, 2025.

[235] Zhao Z, Jing H, Liu X, Mao J, Jha A, Yang H, Xue R, Zakharor S, Guizilini V, Wang Y. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation [J]. arXiv preprint arXiv:2510.08807, 2025.

[236] Gupta A, Kumar V, Lynch C, Levine S, Hausman K. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning [J]. arXiv preprint arXiv:1910.11956, 2019.

[237] Yu T, Quillen D, He Z, Julian R, Hausman K, Finn C, Levine S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning [C]//Conference on robot learning. PMLR, 2020: 1094-1100.

[238] James S, Ma Z, Arrojo D R, Davison A J. Rlbench: The robot learning benchmark & learning environment [J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3019-3026.

[239] Li C, Zhang R, Wong J, Gokmen C, Srivastava S, Martín-Martín R, Wang C, Levine G, Ai W, Martinez B, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation [J]. arXiv preprint arXiv:2403.09227, 2024.

[240] Mees O, Hermann L, Rosete-Beas E, Burgard W. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks [J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7327-7334.

[241] Liu B, Zhu Y, Gao C, Feng Y, Liu Q, Zhu Y, Stone P. Libero: Benchmarking knowledge transfer for lifelong robot learning [J]. Advances in Neural Information Processing Systems, 2023, 36: 44776-44791.

[242] Zhu Y, Wong J, Mandlekar A, Martín-Martín R, Joshi A, Nasiriany S, Zhu Y. robo-

suite: A modular simulation framework and benchmark for robot learning [J]. arXiv preprint arXiv:2009.12293, 2020.

[243] Nasiriany S, Maddukuri A, Zhang L, Parikh A, Lo A, Joshi A, Mandlekar A, Zhu Y. Robocasa: Large-scale simulation of everyday tasks for generalist robots [J]. arXiv preprint arXiv:2406.02523, 2024.

[244] Tao S, Xiang F, Shukla A, Qin Y, Hinrichsen X, Yuan X, Bao C, Lin X, Liu Y, Chan T k, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai [J]. arXiv preprint arXiv:2410.00425, 2024.

[245] Mu Y, Chen T, Peng S, Chen Z, Gao Z, Zou Y, Lin L, Xie Z, Luo P. Robotwin: Dual-arm robot benchmark with generative digital twins (early version) [J]. arXiv preprint arXiv:2409.02920, 2024.

[246] Zhang S, Xu Z, Liu P, Yu X, Li Y, Gao Q, Fei Z, Yin Z, Wu Z, Jiang Y G, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 11142-11152.

[247] Shukla A, Tao S, Su H. Maniskill-hab: A benchmark for low-level manipulation in home rearrangement tasks [J]. arXiv preprint arXiv:2412.13211, 2024.

[248] Ye S, Jang J, Jeon B, Joo S, Yang J, Peng B, Mandlekar A, Tan R, Chao Y W, Lin B Y, et al. Latent action pretraining from videos [J]. arXiv preprint arXiv:2410.11758, 2024.

编写人员贡献

白皮书的总体组成员包括：蒋树强（中国科学院计算技术研究所）、卢策吾（上海交通大学）、刘华平（清华大学）、杨易（浙江大学）、陈广（同济大学）、韩义恒（北京工业大学）、赵君峤（同济大学）、蔡盼盼（上海交通大学）。编写组成员还包括：林倞（中山大学）、张兆翔（中国科学院自动化研究所）、姜育刚（复旦大学）、马楠（北京工业大学）、李升波（清华大学）、张伟楠（上海交通大学）、胡瑞珍（深圳大学）、斯白露（北京师范大学）、郑伟诗（中山大学）、徐凯（国防科技大学）、陈建（中国农业大学）、潘佳（香港大学）、谌骅（逐际动力）、张文强（复旦大学）、张伟（山东大学）、王越（浙江大学）、高飞（浙江大学）、高阳（清华大学）、元辉（山东大学）、宋新航（中国科学院计算技术研究所）、穆尧（上海交通大学）、李弘扬（香港大学）、黎向阳（中国科学院计算技术研究所）、金一（北京交通大学）、黄远（中核集团）、赵健（西北工业大学）、董豪（北京大学）、瞿三清（同济大学）、卢凡（同济大学）、吴勇（同济大学）、孙传兴（中国电子技术标准化研究院）。

其中：蒋树强、宋新航、斯白露、胡瑞珍参与了第一章的撰写，宋新航撰写 1.1 节，斯白露撰写 1.2 节，胡瑞珍撰写 1.3 节；张伟楠、穆尧、郑伟诗、王越、黎向阳、马楠、高飞、张文强、林倞、高阳、姜育刚、李升波、韩义恒参与了第二章的撰写，张伟楠撰写 2.1 节，阚美娜撰写 2.2 节，穆尧撰写 2.3 节，郑伟诗、王越撰写 2.4 节，黎向阳撰写 2.5 节，马楠、韩义恒撰写 2.6 节，高飞撰写 2.7 节，张文强撰写 2.8 节，高阳、林倞撰写 2.9 节，李升波、姜育刚撰写 2.10 节；李弘扬、李升波、赵君峤参与了第三章的撰写，李弘扬、赵君峤撰写 3.1 节，李升波、赵君峤撰写 3.2 节；元辉、潘佳、张巍、徐凯、黄远、陈建、赵健、金一、张文强参与了第四章的撰写，潘佳、谌骅撰写 4.1 节，徐凯、黄远撰写 4.2 节，陈建撰写 4.3 节，金一撰写 4.4 节，张文强撰写 4.5 节；张兆翔、张伟、董豪、孙传兴参与了第五章的撰写，张伟撰写 5.1 节，董豪撰写 5.2 和 5.3 节，孙传兴撰写 5.4 节；陈广、瞿三清、卢凡、吴勇参与了第六章的撰写。蒋树强、卢策吾、刘华平、杨易、陈广、蔡盼盼、赵君峤、韩义恒负责白皮书的框架设计、整体撰写和修订工作。