智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 具身智能的研究与应用

张伟男, 刘挺

引用本文:
张伟男, 刘挺. 具身智能的研究与应用[J]. 智能系统学报, 2025, 20(1): 255-262.
ZHANG Weinan, LIU Ting. Research and application of embodied intelligence[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(1): 255-262.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202406044

DOI: 10.11992/tis.202406044

# 具身智能的研究与应用

张伟男，刘挺

（哈尔滨工业大学 计算学部，黑龙江 哈尔滨 150001）

**摘　要：**随着深度学习和大模型技术的不断增强，人工智能技术从研究简单、封闭的虚拟场景，发展到研究更为复杂、开放的现实场景。研究焦点也从早期的小规模语料库和网络文本数据集处理，发展到多模态一体化的处理架构和研究范式。与此同时，以 OpenAI Sora 为代表的物理世界近似和仿真模型的出现，标志着人工智能再次向通用人工智能迈进了一步。然而，若要让人工智能真正达到通用人工智能的标准，成为类人的智能，需要当今的人工智能体具备与物理世界交互学习的能力，即具身智能。因此，本文主要关注具身智能的研究内容和进展，具体包括具身感知、具身认知和具身行为优化 3 个方面。同时结合近期人形机器人的发展，概述具身智能技术在人形机器人等载体上的应用，并对未来的研究及应用进行展望。

**关键词：**具身智能；具身感知；具身认知；具身行为优化；深度学习；人工智能；仿真环境；人形机器人

**中图分类号：**TP391; TP242.6　　**文献标志码：**A　　**文章编号：**1673−4785(2025)01−0255−08

**中文引用格式：张伟男，刘挺. 具身智能的研究与应用 [J]. 智能系统学报，2025, 20(1): 255–262.**

**英文引用格式：ZHANG Weinan, LIU Ting. Research and application of embodied intelligence[J]. CAAI transactions on intelligent systems, 2025, 20(1): 255–262.**

# Research and application of embodied intelligence

ZHANG Weinan, LIU Ting

(Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

**Abstract**: With the continuous enhancement of deep learning and large model technologies, artificial intelligence has evolved from studying simple, closed virtual environments to exploring more complex, open real-world scenarios. The research focus has shifted from early small-scale corpus and web text data processing to multimodal integrated processing architectures and research paradigms. Meanwhile, the emergence of physical world approximation and simulation models, represented by OpenAI Sora, marks another step forward in the progress towards general artificial intelligence (AGI). However, for AI to truly reach the standard of AGI and become human-like intelligence, it must possess the ability to interact and learn from the physical world—this is known as embodied intelligence. Therefore, this paper primarily focuses on the research content and progress of embodied intelligence, specifically including embodied perception, embodied cognition, and embodied action optimization. Additionally, considering recent developments in humanoid robots, this paper outlines the application of embodied intelligence technologies in carriers such as humanoid robots and provides an outlook on research and applications in this field.

**Keywords**: embodied AI; embodied perception; embodied cognition; embodied action optimization; deep learning; artificial intelligence; simulation environment; humanoid robots

目前，学术界和工业界普遍认为人工智能的发展已经经历了运算智能和感知智能两个明确的阶段，其中被人们所广泛熟知的运算智能的典型应用为棋类程序，如 IBM 的深蓝程序曾经战胜了人类国际象棋世界冠军卡斯帕罗夫[1]，DeepMind 研发的 AlphaGo 围棋程序战胜了多位世界顶级围棋选手[2]。运算智能的成功为人工智能技术的公众接受和感知创造了良好的前期条件。但棋类运动仅作为人类生活中的一个简单场景，并不能真正反映类人的通用智能技术的进步。感知智能以声音、图像、文字的识别技术为代表，在模式识别和机器学习技术的基础上，将人工智能技术进一步推向更大的应用场景，在此过程中也产生了大量用于各行业和领域的应用，如手写体识别、语音输入法、人脸识别、图像处理、多媒体检索等技术，并随着互联网时代的发展，得到了广泛的应

**收稿日期：**2024−06−26.

**基金项目：**国家重点研发计划项目（2022YFF0902100）；国家自然科学基金项目（92470205）；黑龙江省自然科学基金项目（YQ2021F006）.

**通信作者：**刘挺. E-mail: tliu@ir.hit.edu.cn.

用，进一步将人工智能推向更广泛的应用场景。

感知智能技术得公众对人工智能技术的发展有了一定的认识，而随着以 OpenAI 的 ChatGPT[3] 为代表的大模型及其应用的出现，人工智能的发展呈现了跨越式的进步，由感知智能阶段直接步入了创作智能阶段（在以 ChatGPT 为代表的大模型出现之前，学术界和工业界普遍认为，在感知智能和创作智能之间，还需经过认知智能和决策（预测）智能的发展阶段），人工智能技术突破性地实现了在更广阔的行业、领域和人群范围内的普及和认知。

然而，大模型技术及应用在一刻不停地发展，并且从建模文本数据的语言大模型，发展到了建模多模态图像、音视频等时间和空间维度的世界模型，其典型代表就是 OpenAI 发布的 Sora 模型[4]。Sora 模型通过遵循 OpenAI 长期坚持的数据与模型性能之间的缩放法则（scaling law），在未显式引入现实世界物理定律的基础上，仅通过在大量数据上的学习，就实现了对现实世界物理现象的模拟和仿真，进一步拉近了虚拟和现实之间的距离，使得人工智能技术有望从封闭的实验数据集上的学习发展为直接与现实世界的交互和探索中学习，为人工智能发展下一阶段的具身智能提供了新的机遇。

在心理学意义上，“具身”的基本含义是指认知对身体的依赖性；非具身则区别于具身，其认为认知仅仅是一种信息的表征与加工，而与承载它的物理载体（即身体）无关[5]。心理学上，又将具身分为了弱具身和强具身，其中弱具身强调认知对身体的依赖性，但又保留了认知的计算和表征功能；强具身则认为认知是被身体作用于世界的活动塑造出来的，即身体造就了认知。值得一提的是，关于具身认知理论上，笛卡尔曾提出身心二元论，认为人的知识来源与身体的感觉经验无关，即身心为分离的两个世界，而梅洛庞蒂则认为认知的主体是身体，身体和认知是相统一的[6]。梅洛庞蒂的思想成为了具身智能研究的哲学基础。

目前具身智能的研究仍处于起步发展阶段，尽管现在还没有一个被普遍认可的定义，但是可以从其他形式的智能和其实现的功能角度来对其进行定义，即具身智能是一种具有物理载体的系统型功能，其以人工智能、机器人、机械制造及设计等为理论和技术基础，通过与真实物理世界的智能化交互完成特定任务，实现对物理世界的直接或间接影响[7]。非具身的智能通常表现为抽象的智能，典型任务包括围棋游戏、文本处理和图像识别等传统的人工智能任务。具身智能则通常表现为依赖物理载体且能够与环境进行交互的智能，典型任务包括通过抓取、操纵和移动等行为实现的感知和认知，如有遮挡的物体识别和物体物理属性的感知等。

具身智能有广泛的应用场景，如面向家居场景的机器人、面向特种服务和物流场景的机器狗和仓储机器人以及面向更大范围场景的工业机器人。在研究上，当前具身智能的研究热点主要包括 3 个方面，即具身感知、具身认知和具身行为优化。

## 1　具身智能研究内容

### 1.1　具身感知

具身感知的主要目标是让机器能够主动理解周围的环境，其中人也属于环境的一部分。因此，具身感知从大的方面可以分为对物体的感知和对人的感知，其中对物体的感知包括对物体外形、物理属性、几何结构的感知[8] 以及与场景中物体的交互感知；对人的感知则包括对人的意图和行为的感知。

#### 1.1.1　对物体的感知

**对物体外形的感知**[9]　即机器通过自身的移动，获取物体多个视角的信息。之后机器通过融合这些多视角信息，并结合常识重建出物体的外形及物体的外部颜色。对物体外形感知的挑战主要是物体被遮挡，或是在移动中观察物体外形时，对物体定位和外形理解的难度急剧增加。

**对物体物理属性的感知**　物体的物理属性通常包括物体各部分的质量、体积、惯性、摩擦系数、软硬程度等。具身感知通过视觉、触觉和力反馈等方式获得相应的传感信号，从而对物体的各类物理属性有所感知[10]。从上述的任务定义上便可得知具身感知与传统的视觉感知任务之间的差异。

**对物体几何结构的感知**　在具身感知的研究中，对物体的几何结构感知主要体现在对物体操作自由度的感知上，即从可操作的角度，感知物体是刚性还是柔性，以及相对应的自由度。如三维立方体形状的物体其自由度包括 3 个方向上的移动和相应的绕坐标轴转动，因此其自由度为 6[11]；铰接物体在 6 自由度的基础上还有 $n$ 个自由度的操作范围；以及柔性物体，如毛巾、衣服等具有无限维度的自由度等。现有的数据集无法穷尽所有物体的关节结构[12]，因此，具身感知的挑战之一是需要智能体能够自主地与从未见过的高自

由度物体进行交互, 并正确感知其结构。

**与场景中物体的交互感知**　相比于传统视觉感知, 具身感知除了会对物体进行更加丰富的物理属性、几何结构等信息感知外, 还可以与环境进行交互, 并在与环境交互的过程中获取更多的信息, 这被称为交互感知。根据交互方式的不同, 与物体的交互感知可以分为: 移动探索的交互感知[13] 和操作物体的交互感知[14]。

移动探索的交互感知是指机器需要通过在环境中进行移动探索并发现物体, 之后才能对物体进行准确感知。操作物体的交互感知是指机器需要对被感知的物体本身或其他关联物体进行操作, 之后才能感知其物理属性或几何结构等。

### 1.1.2　对人的感知

对人的感知包括对人的意图和行为的感知, 通常也叫做对操作语义的感知。操作语义感知需要机器能理解观测到的人或其他智能体的行为[15], 以便进行后续的任务。而机器对人或其他智能体行为的理解需要模型具有广泛的常识和逻辑推理能力。例如, 当救援机器人看到一个人在泳池中举止笨拙, 机器人需要理解并确认人是溺水了还是在学习游泳, 是否需要急救, 以便判断后续是否采取施救动作和措施。再如, 对于一个家庭服务机器人, 看到家中的人员打了喷嚏, 是否询问其健康状况或者直接递去纸巾等。感知人的意图和行为对于具身智能的研究和应用十分重要, 是具身感知中的重要组成部分。

## 1.2　具身认知

具身认知的目标是对齐虚拟世界与现实世界, 其认知的是与环境交互所需要的能力。通常来说, 非具身的认知任务包括理解抽象的文本、图像或视频等信息, 并输出文本、图片或视频等抽象信息到真实世界中。非具身的认知仅能通过间接的方式与环境交互; 而具身的认知则理解来自真实环境或仿真环境的交互数据, 学习如何与环境进行交互, 包括在仿真环境中学习交互所需要的规划能力与执行能力, 学习完成后则可以在真实环境中与环境进行交互。例如, 人和机器人说: “请帮我拿一枝花。”具备非具身认知的机器人的回应结果可能是一句话: “好的, 给您一枝花。”而具备具身认知的机器人的回应则是将人说的话进行意图理解, 并执行人的指令, 如果通过对环境的探索找到了一枝花, 则将这支实物的花交给人, 否则回复没有找到花。

### 1.2.1　具身认知的核心任务

具身认知的核心任务是对人类指令的理解与执行。以人学习游泳为例, 抽象语言指令可以是“如何游泳？”, 对于具身认知来说, 该指令能够分解成可执行的子任务, 如漂浮、摆腿、摆臂、换气等。在具体执行的过程中, 每个子任务又需要分解成相应的技能, 如对肌肉、关节、呼吸系统的控制等。因此, 机器人的具身认知过程对应可分为

1) 任务规划[16]: 理解抽象指令, 将其分解为可执行的子任务。

2) 技能学习[17]: 学习如何利用已具备的技能执行各项子任务。

而在执行具体技能时, 可以借助工具学习使大模型学会调用机器人操作系统中的底层应用程序编程接口(application programming interface, API)完成执行动作。图 1 所示为具身认知任务规划及执行的示例。

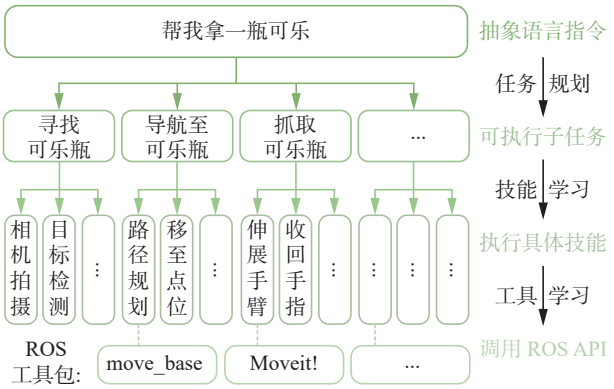


**图 1　具身认知任务规划及执行示例**

**Fig. 1　Example of embodied cognition task planning and execution**

其中, “move_base”与“ Moveit!”为成熟的机器人操作系统 (robot operating system, ROS) 工具包, 机器人可通过调用这些工具包准确完成例如移动、伸展手臂等动作。具身认知模型将“帮我拿一瓶可乐”这种抽象的语言指令通过任务规划、技能学习和工具学习技术, 层级式地拆解为子任务、具体技能和 ROS 工具包, 从而为机器人理解抽象语言指令并准确执行提供了可能。

### 1.2.2　具身认知技术的发展阶段

具身认知技术的发展经历了 3 个阶段, 即规则阶段、拟合阶段和创造阶段。仍以人学习游泳为例: 规则阶段的具身认知技术主要依靠或严格遵照游泳手册中的规则说明来进行类似专家系统的学习, 从而执行具身认知任务。拟合阶段则加入了专家的经验, 该经验可以是来自于真人的示教视频, 通过视频学习游泳相关的技能, 或是通过教练进行现场指导学习游泳。创造阶段可以类比为一个运动天赋优异或在其他运动上有较好经验和技能的人, 将其在其他运动上的经验泛化到

游泳的学习中。图 2 为具身认知技术发展阶段的示意，其中智能性主要指泛化性、适应性和自主性。

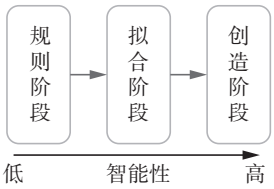


**图 2　具身认知技术的 3 个发展阶段**

**Fig. 2　Three stages of development of embodied cognition technology**

**规则阶段的具身认知**　处于规则阶段的具身认知方法的特点为认知的过程是按照预设程序来处理任务与实现技能的。仍以人学习游泳技能为例，如果按照规则的方式处理，则人需要遵循设定好的时间间隔进行换气，以及按照固定的角度进行摆臂运动，严格执行相应的规则。同理，对于机器来说，其所需要执行的每个特定任务，都由人来提供一套操作或执行步骤，对于新的任务则需要由人来规划新的步骤，并且在执行的过程中步骤不能改变。因此，基于规则方式的具身认知过程，存在着灵活性和泛化性不足的问题，所应用的场景往往是固定的模式和程序步骤，如生产线上的装配机器人。

**拟合阶段的具身认知**　拟合阶段的具身认知技术主要基于对专家示范的简单模仿，使用专家数据训练模型[16-17]，尽可能减少模型输出与专家标签之间的差距。专家数据的来源主要有两个方面，一是机器人的视频[18-19]，即录制的人类遥控机器人完成某些任务的视频。二是人类的视频[20]，即采用互联网海量的人类动作或完成任务的操作视频。如果以人学习游泳为例，拟合阶段的具身认知相当于看着游泳的示教视频来学习游泳。

**创造阶段的具身认知**　创造阶段的具身认知技术得益于大模型强大的泛化能力，借助大模型拥有的丰富世界知识，在跨领域、跨任务上进行泛化的智能决策[21-23]。在任务规划方面，大模型强大的逻辑推理能力能够提高具身模型的任务分解能力，从而将抽象的语言指令拆分为合理的、可执行的子任务[21]。在技能学习方面，大模型强大的语言理解能力能够加深具身模型对技能语义的理解，并通过为其设计奖励函数[22]，增强其对技能的掌握。

综上所述，具身认知的 3 个阶段在任务规划与技能学习的实现方法、适应性、自主性、数据依赖性、应用场景等方面均有显著差异，表 1 以对比的方式对这些差异进行了详细展示。

**表 1　具身认知技术的 3 个发展阶段对比**

**Table 1　Comparison of three stages of development of embodied cognition techniques**

| 特征/阶段 | 规则驱动 | 基于专家示范的拟合 | 大模型驱动的创造 |
|---|---|---|---|
| 定义与特点 | 依赖预设规则和程序 | 通过模仿专家示范训练模型 | 利用大模型能力 |
| 任务规划与技能学习 | 由人类拆解和编程 | 模型从专家数据学习 | 大模型辅助的自主学习 |
| 适应性和自主性 | 缺乏适应性和自主性 | 适应性有限，自主性较差 | 显著提高适应性和自主性 |
| 局限性 | 难以应对变化任务 | 泛化能力受专家数据限制 | 需要大数据和计算资源 |
| 数据依赖性 | 低 | 高 | 极高 |
| 更新和迭代能力 | 困难 | 中等 | 较强 |
| 理解新任务能力 | 低 | 中等 | 高 |
| 实现复杂度 | 低 | 中等 | 高 |
| 应用场景 | 固定环境 | 提供大量示范任务 | 灵活变化的环境 |

### 1.3　具身行为优化

具身行为优化研究的动机是解决模型训练所使用的仿真环境与模型测试或应用的真实环境之间的差距问题[24-26]，从而让模型在真实场景中能够准确地按照预期的目标执行相应的技能。

具身行为优化主要面临 3 个方面的挑战。首先是仿真环境与真实环境的差距，即仿真环境物理引擎无法完全拟真[27]，导致仿真环境中学习的技能在真实物理环境中执行失败。例如：物理引擎难以将物体运动过程中的摩擦阻力、空气阻力等各种因素都模拟的与真实世界完全一致。

其次是仿真环境可学习数据数量与真实环境的差距，即现实环境中天然存在大量有标注数据，而仿真环境中并不存在，需人为构建。数据收集的困难导致仿真环境中机器人的学习效果不理想[28]。例如：现实环境存在大量含有图片和文字的儿童动物绘本，购物网站包含大量有图片和文字描述的商品。

最后是机器人技能学习策略与人类偏好的差距，即机器人在仿真环境中学习技能的策略与真实人类偏好存在差异。例如：机器人学习策略认为水中学习蛙泳最方便易用，但人类普遍认为蛙

泳不够美观而选择学习自由泳。

## 2　具身智能研究的机遇及挑战

### 2.1　具身智能研究的机遇

大模型的出现能够在多个方面助力具身智能技术的发展, 其中包括:

1) 感知更加细腻: 大模型帮助更高效地理解物理世界和人类世界。

2) 决策更加智能: 使用大模型的常识与逻辑推理能力增强机器人决策效果。

3) 执行更加精准: 大模型能够辅助生成训练数据、评价技能任务是否完成[21]。

4) 落地更加现实: 大模型能够在抽象智能以及对人类世界的理解上提供强大助力。

同时, 在大模型之外, 具身智能研究需要新的模型与算法, 通过交互提升机器的感知、认知和决策能力, 即通过与真实世界进行交互而学习新知识。机器人为各种感知、认知、决策算法提供了落地平台, 研究人员可以在真实环境中对算法进行测试, 从而促进实验平台的发展[29]。具身智能是人工智能、机器人学、人机交互等多学科融合的研究方向, 具备多学科交叉属性, 能够从不同学科角度出发进行创新研究。最后, 具身智能的研究可以促进技术的转移, 推动其他研究方向的发展, 如计算机视觉的研究进展为机器人学提供了新的研究思路和方法、大模型促进了多模态一体化表示及处理的新范式等。

### 2.2　具身智能研究的挑战

尽管人工智能(尤其是大模型)、机器人学以及智能制造的发展, 为具身智能带来了前所未有的机遇, 但同时也带来了多方面的挑战。

1) 机器人知识储备的挑战: 具身感知的本质是理解世界, 前提是要具备丰富的世界知识和经验性知识。

2) 机器人复杂逻辑推理的挑战: 具身认知依赖模型的逻辑推理能力, 但如今的人工智能大模型还不具备具身智能所需要的复杂逻辑推理能力。

3) 机器人现实部署的挑战: 由于具身行为优化技术仍未能解决机器人如何在现实中准确执行命令和稳定运行的难题, 导致机器人在现实部署的难度极高。

4) 机器人持续学习进化的挑战: 人类社会在发展, 机器人也要不断地学习新工具、提高自身能力[30], 同时保证学而不忘。

5) 机器人量产和商业化的挑战: 智能化算法需要达到低资源、低成本、高可控性、高稳定性的商业化、产品化需求。

## 3　具身智能的应用

### 3.1　具身智能的典型应用: 人形机器人

哈尔滨工业大学刘宏院士对人形机器人的定义是具有仿人外形、适应复杂环境、执行多任务作业的一种通用机器人, 是机器人研究领域顶端的明珠。

人形机器人的技术和应用, 经历了较长时间的发展。1972 年, 日本早稻田大学加藤实验室研发了世界第一台全尺寸人形机器人 WABOT-1[31], 但其行走一步需要 45 s, 步伐也只有 10 cm, 因此, 其运动能力存在较大的提升空间。2000 年, 日本本田公司制造的 ASIMO[32] 历经数次迭代, 掌握了双足奔跑、搬运托盘、上下楼梯等功能, 标志着人形机器人在运动能力方面的重大进步, 但是此时仍然没有清晰明确的应用场景。2008 年, 法国公司 Aldebaran 研发的小型教学陪伴用人形机器人 NAO, 在教育场景实现了商业落地, 推动了人形机器人的产业化发展。在运动能力方面最为人熟知的人形机器人是美国波士顿动力公司在 2013 年发布的 Atlas 机器人。时至今日, Atlas 已完成了由液压到电控驱动的转型。

如果说之前的人形机器人的发展关注重点在运动控制方面, 那么当前新一轮的发展则是由于以智能化为驱动力而引发的研发上的不断持续投入和应用场景的广泛关注。Tesla 研发的 Optimus 人形机器人, 将汽车智能驾驶的视觉处理系统引入人形机器人, 提升人形机器人的感知及操作能力。通过 OpenAI 的技术加持, Figure 01 机器人具备了强大的与人深入交流, 同时独立做出决策和执行命令的能力。具体而言, Figure 01 通过 OpenAI 提供的多模态大模型获得视觉和文本的理解能力, 并通过策略神经网络获得快速、低级、灵巧的机器人动作执行能力。斯坦福团队开发的 Aloha 低成本全身远程操控系统[33] 可用于机器人操作的数据收集和模拟学习训练, 该系统搭配此团队开发的协同训练算法, 可实现人类对每个任务仅演示 50 次, 即可使机器完成该任务的成功率达到 90% 以上, 且可完成复杂任务。我国乐聚公司研发的人形机器人夸父结合了多模态大模型和模仿学习的优势技术, 在智能化、运动性能等方面均有显著提升, 并在教育和生活场景进行了应用。宇树科技发布的多款人形机器人和机器狗等产品远销海外, 对具身智能和人形机器人的研究和应用, 起到了重要推动作用。

综上所述，人形机器人的新一轮研究和产业热潮是由以大模型为首的人工智能突破性进展导致的智能化程度提升引发的。而智能化通常体现在大脑和小脑的协同控制方面，其中大脑（智能控制）的主要作用体现在对环境感知理解、行为智能控制、自然人机交互、任务决策规划，而小脑（运动控制）的主要作用则在于多体动力学建模、全身协同运动控制、灵巧操作、导航规划。除了智能化之外，人形机器人还需有结实的本体和灵活的四肢。而人形机器人的应用发展，经历了高集成阶段和高动态阶段，正在向高灵巧、高智能阶段发展。

### 3.2　具身智能应用的机遇及挑战

当前，具身智能在多个场景上展现出了应用的潜力。如在服务场景中，半开放场景下替代人工劳动，比如前台接待员、餐厅综合服务员、便利店服务员等。在工业与物流场景中，非标准化的柔性生产成为可能，例如灵活上下料、汽车产线灵活工作、仓储操作等。在特种工作和军事领域中，可以替代人类完成复杂任务，例如机器狗巡检、机器狗战士等。在家庭生活场景中，随着通用人形机器人的成熟，大部分家务可以由人形机器人完成。在情感陪伴场景中，随着大模型的通识理解能力的提升，智能玩具、智能宠物、智能陪伴机器人等成为可能。以上丰富场景的应用数据积累和迭代开发，都为具身智能的应用提供了良好的发展机遇。

然而，具身智能的应用也同样面临着 4 个方面的问题。

1) 技术成熟度有待提升：无论是大脑智能还是小脑的运动控制，在真实场景中的准确性和稳定性是对产业化应用的挑战[27, 34-36]。

2) 场景定制化开发的局限性：大小脑的技术发展程度有先有后，在跨场景、跨任务的通用技术问题解决之前，仍然需要基于场景进行定制化开发[37]。

3) 缺乏大规模真实数据：在各种场景和领域中，缺少数据积累，导致基于数据驱动的学习方法不能直接迁移到具身智能的研究当中[21]。

4) 成本控制问题：当前机器人本体的成本仍然居高不下，要想真正实现机器人的普及，其成本需要快速下降到商业可行的状态。

## 4　总结及展望

具身智能是人工智能与现实环境交互的重要形态，是从用简单明确问题构造数据集、封闭的训练及测试环境、静态的模型更新机制的传统人工智能技术，发展到复杂开放环境下的交互式学习新范式，为人工智能领域的研究和应用提供了大量的新问题和新挑战。具身智能从机器人的感知、认知、行为优化三大方面切入，通过研究能让机器人感知世界的具身感知、让机器人理解世界并像人类一样活动的具身认知、让机器人能够脱离仿真迈向真实世界的具身行为优化，使机器人能够真正获得类人的智能。与此同时，具身智能所研究的三大方面又能相互结合、相互促进。例如高效的感知能够更好地服务于具身认知，增强机器人对环境事务的理解；优异的具身认知、具身行为优化技术又能为具身感知提供更高效准确的世界交互和环境反馈，进而促进交互感知的性能提升。这三大方面间的紧密耦合，促使具身智能不断迈向新的发展阶段，同时也形成了系统级应用，如人形机器人。

虽然人形机器人技术经历了长时间的发展，但前期主要的技术难点在于运动控制方面。而随着基于强大泛化能力的大模型出现，具身智能及人形机器人的发展迎来智能化发展的热潮，一方面，人形机器人成为了大模型落地应用的重要载体，另一方面，大模型也为人形机器人的通用性，提供了“大脑”支持。

由于具身智能是一个多学科交叉的领域，必将带动多个学科的共同发展和相应产业链条上的技术和应用落地。同时，也因多学科的交叉而在研究上带来更多的机遇与挑战。

## 参考文献：


[1] CAMPBELL M, HOANE J. Deep blue[J]. Artificial intelligence, 2002, 134(1/2): 57−83.

[2] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484−489.

[3] OpenAI. Introducing ChatGPT[EB/OL]. (2022−11−30)[2024−06−13]. https://openai.com/index/chatgpt.

[4] OpenAI. Video generation models as world simulators[EB/OL]. (2024−02−15)[2024−06−13]. https://openai.com/index/video-generation-models-as-world-simulators.

[5] 唐佩佩, 叶浩生. 作为主体的身体: 从无身认知到具身认知[J]. 心理研究, 2012, 5(3): 3−8.
TANG Peipei, YE Haosheng. Body as the subject: from the disembodied cognition to embodied cognition[J]. Psychological research, 2012, 5(3): 3−8.

[6] 叶浩生. 认知与身体: 理论心理学的视角[J]. 心理学报, 2013, 45(4): 481−488.
YE Haosheng. Cognition and body: a perspective from

theoretical psychology[J]. Acta psychologica sinica, 2013, 45(4): 481−488.

[7] 卢策吾, 王鹤. 具身智能 (embodied artificial intelligence)[EB/OL]. (2023−07−22)[2024−06−13]. https://www.ccf.org.cn/Media_list/gzwyh/jsjsysdwyh/2023−07−22/794317.shtml.
LU Cewu, WANG He. Embodied AI(embodied artificial intelligence)[EB/OL]. (2023−07−22)[2024−06−13]. https://www.ccf.org.cn/Media_list/gzwyh/jsjsysdwyh/2023−07−22/794317.shtml.

[8] ZHONG Licheng, YANG Lixin, LI Kailin, et al. ColorNeuS: reconstructing neural implicit surfaces with color [EB/OL]. (2023−12−19)[2024−06−13]. https://arxiv.org/abs/2308.06962v2.

[9] LI Kailin, YANG Lixin, ZHEN Haoyu, et al. Chord: category-level hand-held object reconstruction *via* shape deformation[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 9410−9420.

[10] XU Wenqiang, YU Zhenjun, XUE Han, et al. Visual-tactile sensing for in-hand object reconstruction[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 8803−8812.

[11] LIU Liu, XUE Han, XU Wenqiang, et al. Toward real-world category-level articulation pose estimation[J]. IEEE transactions on image processing, 2022, 31: 1072−1083.

[12] XUE Han, XU Wenqiang, ZHANG Jieyi, et al. GarmentTracking: category-level garment pose tracking[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 21233−21242.

[13] RAMAKRISHNAN S K, JAYARAMAN D, GRAUMAN K. An exploration of embodied visual exploration[J]. International journal of computer vision, 2021, 129(5): 1616−1649.

[14] LYU Jun, YU Qiaojun, SHAO Lin, et al. SAGCI-system: towards sample-efficient, generalizable, compositional, and incremental robot learning[C]//2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022: 98−105.

[15] LI Yonglu, LIU Xinpeng, WU Xiaoqian, et al. HAKE: a knowledge engine foundation for human activity understanding[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(7): 8494−8506.

[16] MEES O, HERMANN L, ROSETE-BEAS E, et al. 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.

[17] MENDEZ-MENDEZ J, KAELBLING L P, LOZANO-PéREZ T. Embodied lifelong learning for task and motion planning[C]//Conference on Robot Learning. New York: PMLR, 2023: 2134−2150.

[18] JIANG Yunfan, GUPTA A, ZHANG Zichen, et al. VIMA: robot manipulation with multimodal prompts[C]//Proceedings of the 40th International Conference on Machine Learning. New York: PMLR, 2023: 14975−15022.

[19] AHN M, BROHAN A, BROWN N, et al. Do As I can, not As I say: grounding language in robotic affordances [EB/OL]. (2022−04−04)[2024−06−13]. https://arxiv.org/abs/2204.01691v2.

[20] DAMEN Dima, DOUGHTY H, FARINELLA G M, et al. Scaling egocentric vision: the ëquation missing\"dataset [C]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 753−771.

[21] DRIESS D, XIA Fei, SAJJADI M S M, et al. PaLM-E: an embodied multimodal language model[C]//International Conference on Machine Learning. New York: PMLR, 2023: 8469−8488.

[22] MA Y J, LIANG W, WANG Guanzhi, et al. Eureka: human-level reward design *via* coding large language models[EB/OL]. (2023−10−19)[2024−06−13]. https://arxiv.org/abs/2310.12931v2.

[23] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: vision-language-action models transfer web knowledge to robotic control[C]// Proceedings of the 7th Conference on Robot Learning. New York: PMLR, 2023: 2165−2183.

[24] JAMES S, WOHLHART P, KALAKRISHNAN M, et al. Sim-to-real via sim-to-sim: data-efficient robotic grasping via randomized-to-canonical adaptation networks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12619−12629.

[25] ZAGAL J C, RUIZ-DEL-SOLAR J. Combining simulation and reality in evolutionary robotics[J]. Journal of intelligent and robotic systems, 2007, 50(1): 19−39.

[26] KADIAN A, TRUONG J, GOKASLAN A, et al. Sim2Real predictivity: does evaluation in simulation predict real-world performance?[J]. IEEE robotics and automation letters, 2020, 5(4): 6670−6677.

[27] FANG Haoshu, WANG Chenxi, FANG Hongjie, et al. AnyGrasp: robust and efficient grasp perception in spatial and temporal domains[J]. IEEE transactions on robotics, 2023, 39(5): 3929−3945.

[28] ONGGO B S S, HILL J. Data identification and data collection methods in simulation: a case study at ORH Ltd[J]. Journal of simulation, 2014, 8(3): 195−205.

[29] DEITKE M, HAN W, HERRASTI A, et al. RoboTHOR: an open simulation-to-real embodied AI platform[C]//2020 IEEE/CVF Conference on Computer Vision and

Pattern Recognition. Seattle: IEEE, 2020: 3164−3174.

[30] KROEMER O, NIEKUM S, KONIDARIS G. A review of robot learning for manipulation: challenges, representations, and algorithms[J]. Journal of machine learning research, 2021, 22(1): 1395−1476.

[31] FU Zipeng, ZHAO T Z, FINN C. Mobile ALOHA: learning bimanual mobile manipulation with low-cost whole-body teleoperation[EB/OL]. (2024−01−04)[2024−06−13]. https://arxiv.org/abs/2401.02117v1.

[32] ZHANG Gu, FANG Haoshu, FANG Hongjie, et al. Flexible handover with real-time robust dynamic grasp trajectory generation[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. Detroit: IEEE, 2023: 3192−3199.

[33] LI Shoujie, YU Haixin, DING Wenbo, et al. Visual-tactile fusion for transparent object grasping in complex backgrounds[J]. IEEE transactions on robotics, 2023, 39(5): 3838−3856.

[34] SHEN Bokui, XIA Fei, LI Chengshu, et al. IGibson 1.0: a simulation environment for interactive tasks in large realistic scenes[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague: IEEE, 2021: 7520−7527.

[35] JING Mingxuan, MA Xiaojian, HUANG Wenbing, et al. Task transfer by preference-based cost learning[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2019: 2471−2478.

[36] KATO I, OHTERU S, KOBAYASHI H, et al. Information-power machine with senses and limbs[M]//On Theory and Practice of Robots and Manipulators. Vienna: Springer Vienna, 1974: 11−24.

[37] KUSUDA Y. The humanoid robot scene in Japan[J]. Industrial robot, 2002, 29(5): 412−419.

## 作者简介：

张伟男，长聘教授，博士生导师，哈尔滨工业大学人工智能学院执行院长兼计算学部副主任，黑龙江省中文信息处理重点实验室副主任，中国计算机学会（CCF）理事、CCF 哈尔滨分部主席，中国中文信息学会社交媒体处理专委会社交机器人专业组组长，自然语言处理领域顶级国际会议（CCF A 类）ACL Dialogue and Interactive Systems 资深领域主席。主要研究方向为人工智能、大模型、具身智能、社交机器人。2016 年获黑龙江省科技进步一等奖，2020 年荣吴文俊人工智能科学科技进步二等奖，2022 年获黑龙江省青年科技奖，2024 年获黑龙江省科技进步一等奖。主持国家重点研发计划青年科学家项目、国家自然科学基金面上项目，参与科技创新—2030“新一代人工智能”重大项目、国家自然基金重点项目等多项国家、省部级项目。E-mail：wnzhang@ir.hit.edu.cn。

刘挺，长聘教授，博士生导师，哈尔滨工业大学副校长，国家高层次人次，黑龙江省政协教科卫体委员会副主任。工信部高新司“智能机器人”专家组专家，工信部电子信息科学技术委员会信息服务组副组长，教育部人工智能科技创新专家组成员。国家人工智能产教融合创新平台负责人。中国计算机学会会士、中国中文信息学会副理事长，黑龙江省“人工智能”头雁团队带头人。主要研究方向为人工智能、自然语言处理、具身智能。曾主持国家重点研发计划项目、国家重点基础研究发展计划、基金重点项目。获国家科技进步二等奖（排名第 4）、省科技进步一等奖（排名第 1）2 项，吴文俊人工智能科技进步奖二等奖（排名第 2）。以第一作者出版教材及译著 4 部。E-mail：tliu@ir.hit.edu.cn。