DOI 10.15302/J-SSCAE-2025.07.019

具身智能发展趋势与展望

郑南宁*，杨勐，姜维周，孙宏滨，丁宁

（西安交通大学人工智能与机器人研究所，西安 710049）

摘要：人工智能的发展目标是使机器像人类一样思维和行动，不仅能求解复杂问题，更重要的是能在一个复杂、动态、不确定的物理世界中进行交互。具身智能强调智能体通过物理载体与环境的动态交互，在感知、决策与行动中不断学习和进化，从而突破传统静态数据训练模型的局限，展现出更强的环境适应性与泛化能力，已成为实现人工智能发展目标的关键路径之一。本文深入探讨了具身智能的概念、内涵、计算框架与系统实现，在此基础上进一步梳理了具身智能的发展现状、演进趋势与面临的挑战。同时，特别指出，生成式人工智能，尤其是大语言模型、多模态大模型以及正在演进的“信息–物理–认知”三域融合大模型等技术在加速具身智能演进中的关键作用。面对全球人工智能竞争日益加剧的态势，总结与分析了我国在具身智能领域发展取得的进展和面临的风险，并提出了我国应重点布局的研究方向和针对性的对策建议，助力我国在全球具身智能竞赛中占据领先地位。

关键词：具身智能；人工智能；生成式人工智能；环境交互

中图分类号：TP18；TP24　　**文献标识码：**A

Embodied Intelligence: Development Trends and Prospects

Zheng Nanning*, Yang Meng, Jiang Weizhou, Sun Hongbin, Ding Ning

(Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, Xi’an 710049, China)

Abstract: The development goal of artificial intelligence is to enable machines to think and act like humans, solving complex problems, and more importantly, interacting effectively in a complex, dynamic, and uncertain physical world. Embodied intelligence emphasizes that intelligent agents continuously learn and evolve from perception, decision-making, and action processes. It is realized through dynamic interactions with their surroundings via physical embodiments. This approach overcomes the limitations of traditional static data-driven training models, demonstrating superior adaptability and generalization capabilities in the real world. It therefore has become a dominant way to achieve the goal of artificial intelligence. This study explores the conceptual connotations, computational frameworks, and system implementations of embodied intelligence, and, on this basis, further reviews its current development status, evolutionary trends, and challenges. In particular, the study highlights the pivotal role of generative artificial intelligence, especially large language models, multimodal large language models, and the advancing large “information – physical – cognitive” models, in accelerating the evolution of embodied intelligence. In the face of intensifying global competition in artificial intelligence, this study further summarizes the achievements and analyzes the risks in the development of embodied intelligence in China, and proposes key research directions and targeted policy recommendations to help China secure a leading position in the global

收稿日期：2025-07-10；**修回日期：**2025-08-16

通讯作者：*郑南宁，西安交通大学人工智能与机器人研究所教授，中国工程院院士，研究方向为模式识别与智能系统；
E-mail：nnzheng@xjtu.edu.cn

资助项目：中国工程院咨询项目“生成式AI与具身智能发展战略及对策研究”(2024-XZ-14)

本刊网址：sscae.engineering.org.cn

race for embodied intelligence.
Keywords: embodied intelligence; artificial intelligence; generative artificial intelligence; environment interaction

一、前言

人工智能（AI）的发展目标是使机器像人类一样思维和行动，不仅能求解复杂问题，更重要的是能在一个复杂、动态、不确定的环境和物理世界中进行交互。传统智能系统主要依赖封闭场景、仿真场景或者互联网收集的数据进行模型训练，这种数据训练方式无法构建与现实世界动态交互的闭环学习机制，导致智能系统往往难以适应真实的物理世界。具身智能是一种基于物理实体对环境进行感知与适应性交互，进而理解问题、产生智能行为的智能系统，可以突破传统智能系统依赖静态数据表征的局限，是实现AI发展目标的关键路径之一。

具身智能的概念是AI先驱艾伦·图灵在20世纪50年代首次提出的[1]。同一时期，控制论的创立者诺伯特·维纳也提出了类似的行为智能[2]。20世纪80年代，罗德尼·布鲁克斯和罗尔夫·普费弗等学者在此基础上进一步发展了行为主义智能和身体化智能理论[3]，我国科学家在“国家高技术研究发展计划”的“智能机器人主题”战略规划中也提出了物体的识别与行为交互智能。直到近年，随着AI计算模型不断涌现、算力极大提升和数据易获性增强，人类长期以来一直追求的具身智能，即通过物理实体（智能体）与环境的交互，使智能系统具有环境的适应性及其智能行为的进化，才真正成为可能[4]。目前，具身智能正在引领AI发展的前沿，有望在智能制造、智慧城市、人机协作等关键应用场景中实现技术突破与示范引领，其产业发展将带来显著的经济和社会效益，大大提升生产效率，推动社会全面进步。

当前，我国在具身智能领域的技术积累、数据资源、人才培养及市场规模等方面已取得显著进展。面对全球AI竞争日趋激烈的新形势，亟需加快具身智能核心技术的自主攻关与体系化战略布局，推动AI实现跨越式发展，抢占新一轮科技革命和产业变革的制高点，抓住重塑全球竞争格局的战略机遇。本文系统梳理了具身智能的核心概念与计算框架，结合国际发展态势，全面总结我国在该领域的阶段性成果与面临的挑战，并据此提出我国下一步发展应重点布局的研究方向与针对性对策，助力我国在全球具身智能竞赛中占据领先地位。

二、具身智能的概念与实现

（一）定义与内涵

具身智能打破了传统AI将“智能”局限于大脑内部处理的范式，具身智能体能够通过与环境的持续交互，实现信息采集、认知重构与策略演化的闭环过程。这一理念不仅重构了智能系统的结构设计，也为AI在开放环境中实现更高层次的自主性与适应性提供了理论基础与技术路径。未来，具身智能有望在多场景、多任务、多智能体协同中释放出更强的泛化能力和进化潜力，推动AI迈入真正“类人”认知的新阶段。

1. 基本概念

具身智能通过构建具有本体感知与行动能力的智能体，利用多模态传感器实时捕获环境状态，利用执行机构施加物理作用，并在连续时空维度中形成“感知–认知–决策–行动”的闭环学习系统，从而实现对非确定性环境的动态建模与策略优化。“具身”的含义并非单纯指代物理实体，而是与环境交互以及在环境中执行的整体需求和功能[5]。具身智能强调智能体在物理环境中身体与智能的相互依赖，主张智能不仅仅是大脑的产物，还包括身体与环境的互动。其核心观点是，智能行为不仅依赖于内部的信息处理能力，还取决于智能体的感知和行动能力，即通过感知环境并采取适当的行动来解决问题。

2. 具身学习与具身智能

认知根植于身体行动，经验建构于具身交互。从生物进化的角度来看，所有生物的智力活动都依赖于自身身体与环境的交互，通过积累具身经验，不断适应外部环境，从而在行为或行为潜能上产生积极且持久的变化，这一过程被称为具身学习[6]。具体而言，生物体的智能并非孤立存在，而是深受其身体形态及生存环境的影响。认知过程不仅涉及大脑的信息处理，还与物理、生理和心理三个元素相互耦合，形成动态的循环交互。因此，身体不仅

是执行智能任务的工具，更是认知发展的核心组成部分。换言之，智能的演化并非单纯依赖“算法”的优化，而是“身体”与认知过程协同进化的结果。在AI和机器人领域，具身学习这一理念进一步延伸为具身智能，即机器能够自主感知环境、学习、理解并采取适应性行动的能力，如图1所示。通过与环境的持续交互，智能体能够动态调整自身策略，提升决策能力和适应性，从而实现更高级别的智能行为。这种基于身体–环境交互的智能发展模式，需要基于认知科学、机器人学及AI研究的共同发展，以此构建更具适应性和自主性的智能系统。

3. 具身智能与非具身智能

非具身智能方法通常采用“大规模无监督预训练+小样本有监督微调”的范式[7]来训练神经网络，其核心依赖于大量样本和预设的固定模型进行训练和推理。然而，这种学习方法主要基于静态数据分布，可移植性、可扩展性差，只能在约束条件紧、工作对象少的简单环境下工作[8]，难以模拟人类在“大脑–身体”协作下对目标属性的动态感知和发现能力，因此无法实现具备自主进化能力的高级智能。相比之下，具身智能方法可以通过在虚拟环境中训练大模型，以获取常识表征，并在具体应用场景中结合机器学习方法进行模型优化与进化[9]。这一特性使得具身智能在应对复杂、未知、动态变化的场景时，展现出了更强的适应性和进化能力。例如，在物体识别任务中，基于数据与模型驱动的物体识别方法在面对超出训练数据库范围的新目标时，往往难以适应变化，导致识别性能显著下降。这种局限性使得非具身智能在开放环境和未知场景中的应用受限。与之不同，具身智能不仅能够基于交互行为不断调整自身的识别策略，还能通过持续的环境感知和经验积累，动态适应新的目标和场景。

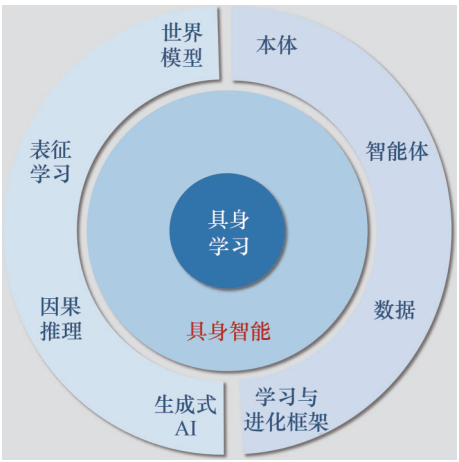

图1　概念分层图

（二）计算框架

当前，具身智能正迈向多技术融合的发展阶段，它的实现依赖于世界模型、表征学习、因果推理和生成式AI等AI理论。世界模型提供环境模拟的结构基座，表征学习提升对信息的抽象与表达，因果推理实现从经验到理解的跃升，而生成式AI则构建起智能体与人类意图及动态环境的统一交互接口，其关系如图2所示。

1. 世界模型：构建认知框架

世界模型[10]用于模拟和预测真实世界的运行规律，通过对物理、社会等环境特征、要素关系的抽象建模，构建出可表征环境动态变化的虚拟系统，无论是视觉场景、物理规则，还是人类行为逻辑，都能被编码进模型中，使其具备对未来状态的预测能力，这为具身智能提供了对环境的理解和预测基础，帮助智能体更好地决策和行动。

2. 表征学习：感知信息的语义化处理

表征学习[11]是将原始数据转换为机器可计算的结构化、语义化数据的过程，其目标是通过自动发现和学习数据的有效特征或表示，降低数据复杂度，提升特征区分度，更好地支持后续的分类、预测、决策等机器学习任务。具身智能在与环境交互过程中产生海量原始数据，例如，传感器信号、视觉图像等，经表征学习能转化为可理解的语义特征，助力智能体快速认知环境、识别物体。同时，学习到的表征可优化智能体的决策和行动规划，而

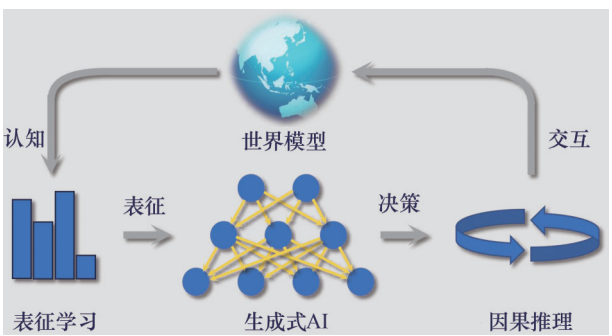

图2　具身智能计算框架

智能体的交互反馈又能不断调整表征学习策略，二者相互作用，推动具身智能高效感知与行动。

3. 因果推理：支撑高阶认知能力

因果推理[12]是基于观测数据或实验干预，在复杂系统中剖析变量之间的因果逻辑，确定指定因素（因）对目标现象（果）产生实际、独立影响的分析过程，通过分析结果变量在其原因变量变化时发生的回应，确定事件或变量之间的因果关系，从而揭示事件背后的因果机制，实现从“关联认知”到“因果理解”的跨越。在具身智能中，因果推理是智能体理解世界的关键能力。智能体分析自身动作与环境反馈之间的因果联系，预判行为后果以优化决策。同时，基于因果推理构建的认知框架，助力智能体快速适应新环境、迁移知识，提升泛化能力。

4. 生成式AI：人机与环境交互的统一接口

近年来，以ChatGPT[13]、SORA[14]、DeepSeek[15]等为代表的生成式AI为具身智能计算框架带来变革。生成式AI[16]是一种基于海量数据和大规模参数的AI技术，能够模拟人类的创造性思维，生成具有一定逻辑性和连贯性的语言、文本、图像、音视频、程序等内容。生成式AI具有强大的理解能力和生成能力，尤其是大语言模型[13]（LLM），融合语言、视觉等多传感器输入的多模态大模型[17]（MLLM），以及正在演进的“信息–物理–认知”三域融合大模型（LIPCM）等利用了超大规模的训练数据并且包含大量参数，使其具备了超强的泛化能力与应用性能，这为具身智能的环境感知提供支撑，同时为具身智能的行动提供决策帮助。

基于生成式AI的具身智能可分为两大部分，即人机交互和系统与环境的交互。在人机交互部分，人以自然语言或图文信息的形式将任务需求输入到多模态大模型中，模型对不同形式的输入进行特征嵌入后，完成任务理解和概念推理，并生成知识和决策，最后由机器人生成面向任务指令的相应行为。在系统与环境交互部分，机器人首先利用自身传感器完成对情境的具身感知，然后根据大模型的学习结果，对情境产生行为，最终完成行为输出。需要指出的是，系统在将情境感知信息输入大模型之前，需要构建一个内部预测模型，在行动之前就能预测到结果。

（三）系统实现

具身智能的系统实现是一个高度协同的工程体系，其包含本体、智能体、数据与学习框架四大核心要素。其中，本体保障执行力，智能体赋予认知力，数据提供驱动力，学习框架实现持续进化。随着模型能力跃升和任务复杂度上升，这一系统正朝向更高效、更泛化、更稳健的方向演进，未来在智慧工业、城市治理、人机协作等领域具备广泛应用前景。

1. 本体：智能落地的物理承载

具身智能的本体指代实际执行物理实体，承担在物理或虚拟环境中进行感知与任务执行的职能，例如四足机器人、复合机器人或人形机器人等。作为连接虚拟世界与物理世界的桥梁，本体需具备环境感知、运动控制与操作执行等基本能力，其能力边界直接制约了智能体任务完成的范围与水平。

2. 智能体：系统的决策与推理中枢

智能体作为物理本体的智慧核心，承担感知、解析、决策与操控等关键职能，理解复杂的环境结构及其语义内容，并与环境动态交互。随着深度学习[18]技术的迅猛发展，当代智能体大多数由深度神经网络模型驱动，特别是语言大模型、MLLM等为智能体提供了更强的环境理解与推理能力。

3. 数据：驱动智能进化的“燃料”

为了广泛适应复杂多变的环境和任务，智能体所依赖的深度神经网络模型的规模正在不断增大，这些模型对于数据的渴求也愈发强烈[19]。对于具身智能来说，场景的复杂性和多样性使得所需处理的环境和任务更加多变，这涉及到围绕复杂任务链的规划、决策和控制数据。特别是，针对特定行业场景的高质量数据，将成为具身智能在未来成功应用和实施的关键支柱。

4. 学习进化框架：实现适应与迁移的机制

学习进化框架通过智能体与物理世界的互动，逐步实现对新环境的适应、对新知识的吸收以及对解决问题能力的不断增强。在初期阶段，利用虚拟仿真环境进行高效学习是一种行之有效的策略，但现实世界的复杂性远超仿真环境。因此，如何实现虚拟与现实环境之间的高效知识迁移，已成为智能体架构设计中不可或缺的一环，该问题的解决将直接影响智能体在真实世界中的表现与适应能力[20]。

三、具身智能的发展态势

（一）发展现状

具身智能正处于基础研究与产业应用双轮驱动的快速发展阶段。在各个关键环节上，核心算法与模型不断突破，智能体感知理解与执行能力持续增强；在机器人形态、仿真平台与典型场景中，多元化的落地探索逐步实现从“能动”向“高效”演进。随着大模型能力下沉、硬件平台成熟和应用需求扩大，具身智能将加速走出实验室，成为引领下一代AI变革的关键力量。

1. 基础研究

具身智能的基础研究围绕“感知–交互–规划–仿真–训练–加速”体系积累了一系列重要成果[21]，其中MLLM与世界模型起到关键作用。

在感知层面，基于主动视觉的感知方式通过语义视觉即时定位与地图构建（SLAM）（如ORB-SLAM[22]）实现了动态环境下的高精度定位与地图构建，结合三维场景理解技术（如Point Transformer[23]等点云处理模型），使智能体能够解析复杂空间关系；具身智能与人形机器人等领域的交叉融合正在形成更加遵循智能机器人以人为本、协作智融的“人本智造”新范式，人形机器人已逐渐成为具身智能的理想载体[24]。此外，具身智能技术将AI融入自主无人系统等物理实体，从视觉、语言等感知中处理信息并进行推理，具备与环境交互的能力[25]。因此，基于机器人触觉的视觉触觉传感器（如GelSight、DIGIT）开始融合多模态数据，在物体重建、抓取规划等任务中显著提升了物理交互精度。

在交互与规划层面，具身问答（EQA）和语言引导抓取成为研究热点。基于LLM的方法通过自然语言解析与环境探索，实现了复杂指令理解；具身智能体架构从SayCan[26]的模块化设计向端到端模型演进，典型模型如谷歌的RT系列：RT-2[27]通过“视觉–语言–动作”模型整合多模态能力，支持“思维链”推理完成工具选择等复杂任务，RT-H[28]则引入动作层次结构优化细粒度控制。视觉–语言–导航[29]及视觉–语言–规划[30]等端到端模型也开始相继出现。然而，当前的LLM或MLLM等专注于单一信息域，未来趋势是向LIPCM发展，将信息域（感知数据）、认知域（推理与理解）和物理域（动作执行）知识统一整合，实现物理环境中的闭环协同。

在仿真层面，仿真到现实技术通过具身世界模型（如Sora[14]等生成模型）模拟物理规律，为具身智能体提供了高效的交互学习环境。然而，现有的大规模仿真平台如AI2-THOR[31]、RoboTHOR[31]及Habitat[32]在复杂物体、动态交互、物理建模上仍显不足。虽然结合域随机化和人类干预校正等技术，缓解了模拟与真实环境的差距，推动模型从虚拟训练到物理部署的迁移，但仿真环境与现实场景之间的域差异仍是亟待解决的问题之一[33]。

在训练层面，构建可复现、可扩展的人形机器人训练场是实现规模化具身智能的关键。训练场可分为：真实试验场（受控场景，如RH20T数据集[34]，便于获取高质量力/触觉/动作数据）、混合现实场（物理+虚拟元素混合，如VinT-6D数据集[35]，用以生成跨域样本）、数字孪生平台（完整场景与部件的虚拟副本如AI2-THOR数据集[31]，支持并行化训练与回放），其多模态数据类型包括：高精度运动捕捉（动作/步态数据）、力/触觉序列（抓取/接触数据）、同步多摄像头与深度感测、触觉与皮肤式传感器数据以及语音与语言交互日志。

在加速层面，具身智能的发展正受到端侧实时性与能效优化的双重驱动。当前具身智能产品在实时性和可靠性要求高的场景，对云端通信的效率和本体侧芯片的推理能力有着更高的要求[36]。当前研究通过非结构化剪枝、知识蒸馏、低秩分解等架构优化与模型压缩方法使智能体在Jetson、昇腾等低功耗平台上也能保持低延迟与高性能运行[37]。未来的发展方向是将元学习和神经架构搜索等自动机器学习技术与模型压缩相结合，在不影响模型性能的情况下进一步提高压缩率。与此同时，如何在特定硬件架构下通过系统层面优化推理的效率成为新的挑战[38]。

2. 应用实践

具身智能的落地实践以机器人为核心载体，在硬件形态、环境模拟与任务执行层面取得了全面突破。在硬件创新层面，固定基机器人如Franka Emika Panda凭借微米级作业精度可以很好地适配实验室自动化与精密制造场景；四足机器人如波士顿动力公司的Spot机器人以高稳定性和地形适应

性，在工业巡检、灾难救援中承担了环境探索任务；人形机器人如特斯拉公司的Optimus则通过类人形态与LLM整合，逐步渗透到了制造业装配、家庭服务等需要人机协作的场景。在环境模拟层面，通用模拟器如Isaac Sim和Gazebo提供了高保真物理引擎与多机器人协同能力，支持算法快速验证；真实场景模拟器如AI2-THOR和Habitat基于三维扫描构建了逼真的室内环境，例如AI2-THOR的RoboTHOR模块包含89个真实公寓场景，可支撑多智能体交互与复杂指令执行研究。

3. 相关产业

具身智能在工业制造、自动驾驶、物流运输、家庭服务、医疗康养等领域都实现了产业落地。根据国际机器人联合会（IFR）的数据，2025年全球工业机器人市场规模预计将突破500亿美元，全球人形机器人市场规模预计也将突破500亿美元。国内外科技企业纷纷布局具身智能，如谷歌公司、特斯拉公司、宇树科技有限公司的人形机器人，Waymo公司、特斯拉公司、比亚迪股份有限公司、百度公司、小马智行科技有限公司的自动驾驶，ABB公司、库卡机器人制造有限公司、新松机器人自动化股份有限公司、北京极智嘉科技股份有限公司等的自动引导车与工业机器人，大疆创新科技有限公司、亿航智能技术有限公司、FreeFly公司等的无人机。2024年8月，第五届中国机器人学术年会在西安举办，具身智能是本次会议展示产品中出现频率最高的关键词。具身智能正在引领AI发展的前沿，具身智能产业发展将带来显著的经济和社会效益，大大提升生产效率，改善人民生活品质，推动社会全面进步。

（二）发展趋势

具身智能正在经历从技术整合到系统落地的关键跃迁期。三域融合大模型实现了感知、认知、执行的闭环支撑，生成式AI与仿真建模加速智能体的训练与演进，轻量化架构确保了边缘部署与响应效率，产业链协同与场景拓展则为应用推广注入持续动力。

1. 单域大模型向“信息–物理–认知”三域融合大模型进化

人对外部世界的认知过程，本质上是一个多传感信息的融合过程，如视觉、触觉、听觉与语言等，且人脑具备抽象逻辑思维能力，通过综合不同模态的信息，形成对世界的更深刻认知[39]。然而当前的生成式AI，如语言大模型、视觉大模型、MLLM等，专注于单一信息域，难以在真实世界中实现感知–认知–执行闭环的动态协同。三域融合大模型整合信息域（感知数据）、认知域（推理与理解）和物理域（动作执行）知识，能够克服单域大模型的局限性，实现复杂开放环境的建模与动态交互，这与具身智能的理念高度契合。从信息处理角度看，三域融合大模型如同人类大脑综合多模态信息一样，将感知数据、推理理解和动作执行知识融合，为具身智能提供了更全面、强大的信息处理基础。从环境交互方面看，三域融合大模型通过对复杂开放环境的建模与动态交互能力，让智能体在感知环境后，能够基于认知域的推理与理解，在物理域精准执行动作，从而实现与环境的有效互动。三域融合大模型将助力具身智能真正在现实世界中发挥作用，推动其从理论走向实际应用。

2. 生成式AI加速与具身智能深度融合

语言大模型、多模态大模型、三域融合大模型等生成式AI技术与具身智能的融合是迈向通用AI的有效途径之一。具身智能注重从环境交互的数据中学习智能决策，使机器具备自主感知环境、学习、理解和行动的能力，从而更好地理解适应复杂动态开放环境。大模型的特点是强大的理解能力和生成能力，因此可以利用大模型的理解能力为具身智能的环境感知提供支撑，同时为具身智能的行动提供决策帮助。二者融合有助于提升机器人在复杂、动态、开放环境中的泛化能力，同时增强具身智能体行为的可解释性与可控性，促进人机协同效率的提升。

3. 仿真环境与世界模型不断完善

智能体通过与虚拟环境、真实环境的交互，实现对新环境的适应、新知识的吸收以及解决问题能力的增强[33]。现有的方法大多利用虚拟仿真环境进行模仿学习。然而真实环境的复杂度往往超越了仿真环境，因此，如何在仿真环境与真实世界之间建立有效衔接，实现高效的知识迁移，是具身智能架构设计中不可或缺的一环。此外，世界模型展示了强大的模拟能力和对物理定律的理解能力，这使具身智能能够全面理解真实环境[40]。通过不断完善仿真环境与世界模型，具身智能将能够更高效地学习

复杂任务并实现在真实场景的泛化。

4. 基础模型架构朝轻量化演进

具身智能应用对部署灵活性和资源效率有着较高要求，这一需求正推动模型架构朝着轻量化方向发展。在未来，具身智能的研究重点之一将在于压缩大模型的参数规模、提高推理效率，以及优化感知与决策模块的架构，以此适配计算资源相对受限的边缘设备与机器人平台。模型轻量化对具身智能发展带来的益处是多方面的：一方面，它能够显著提升系统的响应速度，让智能体对环境变化做出更及时的反应；另一方面，轻量化有效降低了具身智能大规模部署的成本，使相关技术更易于推广应用。通过这些优化，具身智能有望在更广泛的场景中落地，为各领域发展注入新的活力。

5. 产业链上下游协同构筑技术生态

具身智能产业链上下游企业紧密联动。上游供应环节，芯片企业将不断研发更适配具身智能的专用芯片，提升运算速度与能效比；传感器厂商会推出精度更高、响应更快、集成度更强的产品，以满足智能体对环境信息的精准感知需求；材料供应商也将开发出更轻质、高强度且耐用的材料，用于制造智能体的“身体”部件。中游研发制造环节，各大企业会在算法优化上加大投入，结合强化学习、模仿学习等多种技术，提升智能体的决策与学习能力；系统集成商将整合硬件与软件，打造出更稳定、高效的具身智能系统；同时，专业的测试机构会不断完善测试标准与流程，确保产品质量与安全性。下游应用端，各企业将根据自身需求，积极与具身智能企业合作，定制化开发解决方案，加速智能体在实际场景中的落地应用。

6. 多元场景推动实际应用落地

具身智能的应用场景不断拓展。在工业制造领域，具身智能的应用将从简单的搬运、巡检等任务，向更复杂、精细的环节拓展。如在电子产品制造中，能精准完成芯片级别的组装与检测；在汽车生产线上，可实现不同车型零部件的灵活切换与安装，极大提升生产柔性。在医疗与康复领域，具身智能应用有望替代机械式的、普适式的工具，提供更加精准的定制化服务。如在医疗场景中，手术辅助机器人借助具身智能，能依据患者的实时生理数据，辅助医生进行更精准、微创的手术操作；在康复治疗中，智能设备可针对患者的康复进度，定制个性化训练方案并实时调整。在养老、陪护与服务场景中，能够通过自然语言交互、情绪识别与自适应行为提供安全、私密且具情感计算能力的陪护服务，承担接待、引导等交互性更强的工作。在公共安全与灾害响应中，四足机器人、人形机器人与无人机可协同执行灾区侦查和物资投送，依赖异构平台间的实时协同，实现对灾区的全面侦测、勘察、救助及引导等工作。在城市基础设施与巡检领域，通过结合无人机与地面机器人，实现桥梁、管线、输电塔等的高频巡检与主动维护预警，极大地降低维护成本同时提高安全性。在家庭、教育与娱乐方面，低成本、多功能家庭及教育机器人将提供更优的人机交互体验与更高的可靠性和可维护性，承担家务、陪伴老人小孩等多项任务，成为家庭生活的得力助手。

（三）发展挑战

当前具身智能面临的核心问题集中在4个方面：模型不可解释性、算力资源对立性、数据资源稀缺性与生态系统不健全性。这些问题不仅制约了其在现实场景中的稳定运行与规模化推广，也暴露出从技术验证向产业落地过程中尚未解决的系统性挑战。

1. 大模型“黑箱”制约具身智能可控发展

随着语言大模型、多模态大模型等在具身智能系统中的广泛应用，其能力不断增强，但也带来了决策过程不透明、行为难以预测的问题。在具身智能场景中，智能体需在开放物理世界中持续感知、推理并做出实时决策，对模型的可解释性与可控性提出更高要求。然而，当前大模型缺乏对感知、知识与行为之间因果关系的理解机制，容易导致策略偏差、行为异常，增加在复杂真实环境中运行的不确定性与风险，限制了具身智能的可控发展。因此，构建具备世界知识支撑的因果推理机制，是实现具身智能稳健、安全运行的关键前提。

2. 云端训练与边缘部署形成双重技术压力

具身智能在算力层面面临云端训练与边缘部署的双重挑战。一方面，促进具身智能高速发展的大模型的训练高度依赖超大规模计算资源，需使用成千上万张高性能图形处理器（如A100）构建集群，训练周期长、能耗高；另一方面，将模型部署至机器人本体，则需在资源受限的嵌入式平台（如

Jetson等）上实现高效、低延迟的推理和控制，需要高性能的计算芯片，同时亟需依赖剪枝、量化、蒸馏等模型压缩技术。

3. 高质量多模态数据资源严重匮乏

具身智能的发展高度依赖大规模、高质量的多模态数据，包括三维空间信息、传感器数据与运动轨迹等。然而，现有数据资源数量有限、模态单一、质量参差不齐，难以支撑复杂的感知与行为学习。现有开源数据集规模较小，缺乏统一的采集与标注规范，且多为单一模态的数据，限制了模型的跨任务、跨场景泛化能力。同时，国内不同机构与企业间数据封闭，缺乏共享机制，进一步阻碍了规模化、标准化数据体系的建立。

4. 产业生态尚未成熟

具身智能领域具备广阔的产业潜力已成为市场共识，但其发展仍面临产业规模化与市场拓展的制约。一方面，当前软/硬件成本较高，受限于核心技术尚未突破与规模化生产能力不足，难以有效降低产品成本，限制了大规模落地应用；另一方面，市场接受度仍需时间积累，加之实际应用场景、商业模式尚不成熟，进一步影响用户信任与消费意愿。此外，行业标准与监管体系尚不完善，产品性能与安全性参差不齐，制约了具身智能产业的健康可持续发展。

要推动具身智能迈向成熟，必须强化因果推理机制研究、打通云边协同路径、建立高质量数据标准体系、推动产业生态协作机制建设。只有攻克这些基础瓶颈，才能真正释放具身智能的应用潜能，为工业、医疗、服务等领域注入持续创新动力。

四、我国具身智能的发展现状

（一）取得的进展

我国在具身智能领域已形成“政策引导、技术突破、数据支撑、人才集聚、市场牵引”五位一体的发展格局。从国家战略部署到产业落地，从原始数据积累到人才储备，从自主研发到全球竞争力提升，我国具身智能发展已进入加速期。

1. 战略高度重视，政策强力支持

在全球科技竞争的浪潮中，具身智能已成为焦点领域，我国在具身智能领域展现出强劲的发展势头。我国在《2025年国务院政府工作报告》[41]和《“十四五”机器人产业发展规划》[42]等文件中明确提出，培育具身智能等未来产业，将具身智能纳入国家战略部署，集中优势力量攻克具身智能技术难关，提升国家整体科技实力，助力经济高质量发展和国际竞争力提升。在相关政策的引导下，我国在具身智能领域积极布局、大力投入，从技术研发到产业落地，从人才培养到市场拓展，均取得了一系列令人瞩目的成果。不仅在关键技术上实现自主创新突破，还在数据资源、人才储备以及市场应用等方面构筑起独特优势，为具身智能产业的蓬勃发展奠定了坚实基础，正逐步从“跟跑者”转变为“并肩者”乃至部分领域的“领跑者”。

2. 自主成果不断涌现，部分创新技术全球领先

2024年，我国自主开发的多模态AI工具DeepSeek，其最新一代大语言模型DeepSeek-V3[15]在多项评测中表现出色，性能优于广泛使用的ChatGPT，并且在成本上具有显著优势。此外，DeepSeek还发布了R1模型[43]，该模型在技术上实现了重要突破，通过纯深度学习方法让AI自发涌现出推理能力，性能比肩OpenAI公司的o1模型正式版[44]，但训练成本仅约600万美元，远低于美国OpenAI公司的数亿美元投入。DeepSeek的应用程序目前已经超越ChatGPT，其开源模型引发了全球关注，对美国科技行业的竞争力产生了重要影响。2015年，我国AI领域研究人员提出的神经网络基础模型ResNets[45]被广泛应用于ChatGPT、AlphaGo、AlphaFold等重要产品。此外，杭州宇树科技有限公司在四足机器人与人形机器人方面领先全球，其运动控制与硬件设计能力获得世界知识产权组织（WIPO）全球奖项肯定。清华大学开发的“天工”具身智能系统在机器人自主决策与环境交互方面达到国际先进水平。上海智元新创技术有限公司自2024年起开始量产双足类人机器人，其Lingxi X1/Lingxi X2具备高灵活性关节驱动、端到端AI控制系统和丰富的传感能力，并通过Aidea平台促进交互数据大规模开放与共享。深圳众擎机器人（EngineAI）于2024年推出SE01、PM01等人形机器人，其中PM01实现了类人前空翻特技，展现出卓越的运动与平衡能力。

3. 原始数据资源丰富，为具身智能提供动力

数据是具身智能的核心要素与重要资源，我国在数据资源方面具有十分明显的优势，在智能制

造、自动驾驶、智慧城市、医疗健康等具身智能应用领域中不断增加的传感器和智能设备部署，使得我国在工业场景、商业场景以及社会治理场景中能够持续获取海量、多样化、实时更新的数据。例如，在智能制造领域，工业机器人、智能数控机床等设备上的传感器时刻捕捉生产环节的温度、压力、速度等数据，实时反馈产线状态，形成全流程工业数据链。在自动驾驶领域，车辆的激光雷达、摄像头等设备，不仅采集道路环境、交通信号数据，还记录车辆的行驶轨迹、驾驶行为，构建起复杂路况的动态数据库。在智慧城市领域，从智能监控、环境监测站到政务服务终端的多源数据，精准、全面地刻画了城市的日常状态。在医疗健康领域，智能穿戴设备记录用户的生命体征，医疗影像设备产生海量的计算机断层扫描、核磁共振成像数据，为疾病诊断与健康管理提供丰富信息。海量的、多样化的真实原始数据为具身智能算法的训练和迭代提供了充足而多样的“燃料”，驱动其不断进化升级，助力具身智能在复杂现实环境中实现精准感知、高效决策与灵活行动。

4. 人才与企业双轮驱动，创新活力持续增强

统计数据显示，我国过去10年在AI领域的论文数量达2.54万篇，研究型人才数量为1.74万人，均仅次于美国位居全球第二，形成了较强的竞争力，这为具身智能的基础理论研究、算法优化、模型构建等关键技术突破提供支撑，形成了较强的竞争力。此外，工业和信息化部的统计数据显示，我国AI相关的企业数量超过4500家，且数量正在快速增长。其中，众多专注具身智能的创业公司如雨后春笋般涌现，从研发能精准抓取物体的灵巧机械手，到打造具备复杂环境适应能力的人形机器人，不断探索具身智能应用边界。相关研究机构和公司吸引了大量的高端人才，形成了从基础研究到产业应用的完整人才梯队，为我国具身智能的创新和产业化奠定了坚实基础。

5. 内需市场广阔，应用场景持续拓展

根据《2024年具身智能产业发展研究报告》等数据[36]，2024年我国具身智能市场规模已超过8000亿元，并有望在2026年突破万亿规模。政府相继出台相关的扶持政策，吸引了大量社会资本投入，同时我国大量传统行业对自动化、智能化的升级需求，形成了巨大的市场牵引力，推动具身智能技术在制造、物流、医疗、家政等市场的持续增长。依靠卓越的环境适应能力，具身智能已经成为工业制造、国防安全等国家战略领域的核心需求。我国产业基础雄厚，拥有成熟的工业体系、完善的基础设施以及通用支撑平台，并在工业机器人、智能车等领域实现了机器人本体与智能系统的深度耦合，在具身智能应用方面已初步形成优势。

未来，随着基础研究成果进一步积累与产业生态体系进一步成熟，我国有望在具身智能这一未来科技高地上占据更大领先优势，引领全球智能科技新潮流。

（二）面临的风险

虽然我国在具身智能领域已经取得了重要进展，部分核心关键技术实现了重大突破，形成了发展具身智能的良好“土壤”。但同时，也要清醒地认识到，我国具身智能整体发展水平与发达国家相比仍存在明显不足，在具身智能领域的基础模型和新兴计算架构方面缺少重大原创成果，在相关的基础理论、核心算法、集成系统、开源平台等方面差距较大，缺乏战略性超前布局，尖端人才远不能满足需求。

1. 基础研究与系统集成创新能力乏力

我国的具身智能研究具有鲜明的行业特色，虽然在自动驾驶和人形机器人等场景融合转化方面取得了出色的成绩，但在原始创新上还存在差距，尤其在新型神经网络结构设计、基础模型学习理论与方法探索、智能体具身学习理论与方法等方面缺乏长期、持续的研究经费支持，往往以技术成果转化的机制评估基础研究。此外，具身智能研究涉及机械结构、电子器件和AI算法等多个专业领域，虽然国内在部分研究上取得了不错的进展，突破了国外对智能芯片和基础模型的封锁，但是在系统能力上还没有达到整体大于部分之和的集成效果，亟需打破学科壁垒，构建多领域协同创新平台，实现部件间的有机融合，使具身智能产业释放全部潜能。

2. 开源平台建设仍处于起步阶段

具身智能研究高度依赖开源算法和开源数据，开源算法（如ROS、PyBullet等）为研究者提供了标准化的开发工具和仿真环境，大幅降低了核心技术的研发门槛，开源数据（如Matterport3D[46]、Habitat-Matterport 3D[47]等）解决了具身智能训练所

需的物理交互数据匮乏问题。我国目前已发布了“天工开源计划”等开源平台建设计划，开放部分软件开发文档和结构设计文档，但数据集、运动控制训练框架等关键资源的开源仍需进一步完善。同时，训练环境、计算中心等基础支撑平台的建设无法满足迭代研发需要，尚未形成跨学科、跨领域、跨行业的高效协同创新机制。这些问题导致资源分散、重复建设现象严重，难以形成合力，影响了具身智能领域的快速突破和产业化进程。

3. 产业发展长期战略规划仍需完善

具身智能产业发展具有高度交叉、长产业链、强场景依赖等特点，这要求必须构建跨部门协同的政策体系、全链条贯通的标准化框架和“产学研”深度融合的创新生态，才能突破当前发展瓶颈。我国目前相关的政策制定缺乏对长期战略布局的深入思考，一些产业化项目投资少、周期短，还要见效快、效益高，这难以适应具身智能发展的复杂性和不确定性。我国在算法性能、硬件接口、安全规范等方面缺乏统一标准，影响技术迭代和产业协同，相比之下，欧美已建立具身智能的仿真测试标准体系。此外，高校、科研机构与企业间缺乏长效合作机制，基础研究与应用落地脱节，导致具身智能产业发展仍滞后于国际水平。

4. 人才培养与考核机制对技术发展形成制约

具身智能研究融合了AI、机械工程、传感器技术等多学科知识，从基础理论突破到关键技术研发，再到核心器件的迭代优化，往往需要团队协作与长期积累。现有科研人才考核机制往往忽视了具身智能领域研究的复杂性和长期性，导致科研人员往往以“短、平、快”的方式聚焦于短期目标和具体技术突破的现象，对具身智能基础理论和关键技术的研究不够深入。同时，具身智能研究高度依赖跨学科合作，需要不同专业背景的研究人员协同攻关，现有的考核机制难以全面衡量具身智能跨学科研究的成果和贡献，在研究人员跨学科合作中形成了诸多障碍。

我国具身智能发展已初具规模，但要实现全球引领仍面临诸多挑战。必须从基础研究、开源生态、系统集成、产业战略和人才机制等多维度发力，推动多学科协同创新和长期战略部署，加快补齐短板，夯实底座，才能真正释放具身智能对未来科技与产业的引领潜力。

五、我国具身智能下一步发展建议

生成式AI与具身智能正在给AI发展带来颠覆性变革。未来AI领域的竞争，将不再局限于资源与技术的角逐，而更多取决于宏观战略布局与路径选择的智慧。在这一关键领域，DeepSeek率先做出了开创性探索，创新性地采用了“开源模型+闭源产品”的双轨战略。这一战略既通过开源核心模块为全球AI社区注入创新动力，加速打破行业壁垒，促进技术普惠；又通过闭源关键产品，有效保护核心技术优势，确保商业化发展的可持续性。这种战略选择不仅体现了对AI发展规律的深刻洞察，更彰显了开放共赢的发展理念。在这场关乎人类未来的AI战略竞争中，唯有准确把握宏观方向，规避潜在风险，才能掌握发展主动权，最终在激烈的竞争中赢得长远发展。

（一）应重点布局的研究方向

生成式AI与具身智能的深度融合正引发智能系统形态的跃迁，从能力涌现机制到三域融合大模型，从分布式协同架构到轻量化边缘部署，多个关键方向正在共同推动AI系统向泛在化、自主化、协同化演进。构建具备逻辑推理、物理交互和群体智能的未来智能系统，将为我国在具身智能领域实现弯道超车、引领全球提供重要支撑。

1. 生成式AI大模型系统的能力涌现机理

生成式AI大模型的能力涌现源于超大规模参数与多模态数据的非线性交互，通过自注意力机制实现从数据拟合到逻辑推理的跃迁（如GPT-4的思维链能力）。大模型的透明性需结合注意力可视化与因果推理建模，揭示智能决策的依据，同时通过价值观对齐确保生成内容符合伦理规范。

2.“信息–物理–认知”三域融合大模型

通过构建跨域知识表征体系，实现数据空间、物理规律与人类认知范式的深度融合。该方向将突破传统AI的领域边界，使系统具备对复杂场景的涌现式理解能力，在自动驾驶、智能医疗等领域形成“环境感知–物理仿真–决策推理”的闭环智能。

3.“云–端”协同跨域分布式架构

通过动态分层任务分配实现万亿参数模型的分布式训练与推理：云端负责全局知识融合与复杂计算，端侧执行轻量化推理与隐私敏感数据处理，在

保障隐私安全的同时，支持多模态数据的实时处理与知识迁移，为具身智能在智能交通、智慧城市等场景的应用提供弹性扩展的计算架构。

4. 具身智能驱动的多智能体系统

融合生成式AI的创造性与具身智能的物理交互能力，研发支持多异构智能体自主协作的认知架构，推动工业机器人集群的自主协同制造，以及灾害救援等场景中无人机–地面设备的跨模态协作，形成动态环境下的群体智能涌现。

5. 轻量化生成式模型部署

开发基于神经架构搜索的模型压缩技术，实现百亿参数级大模型在嵌入式设备的低功耗运行，通过知识蒸馏技术构建面向服务机器人等终端的轻量化引擎，在保持生成质量的同时将降低推理延迟。

6. 人形机器人智能系统

构建结合大模型驱动的语言理解与物理仿真引擎的具身认知框架，推动养老陪护等场景的机器人实现自然对话、场景自适应操作能力，建立人机共生的新型交互范式。

（二）对策建议

在新一轮科技革命与产业变革浪潮中，AI已成为重塑全球竞争格局的核心力量。我国AI发展必须锚定战略方向，强化核心技术攻关，构建自主可控的创新生态体系。通过系统性布局，在全球AI竞争中赢得主动权，推动我国从AI大国迈向AI强国。

1. 进一步强化国家级科研平台作用

进一步强化国家实验室、全国重点实验室的科技力量，整体规划布局未来具身智能发展方向，加速具身智能相关技术的突破。打造国家级具身智能创新中心，整合高校、科研院所和行业领军企业的优势资源，围绕制造业升级、智慧医疗等国家重大需求开展协同创新。此外，还应积极推动国际合作，推动AI领域多学科交叉研究的国际化，通过与国际知名开源项目的合作，提升国内具身智能研究的影响力，增强我国在全球技术竞争中的话语权。

2. 加强培育基础研究源头创新

促进认知科学、机器人学、神经科学、AI的学科交叉融合，打破学科壁垒。加大资金投入，稳定支持科研项目，营造良好学术环境，鼓励自由探索，优化科学评价体系，以创新成果而非短期效益“论英雄”，要以足够的耐心激发科研人员的积极性与创造力，培养一流科学家和卓越工程师。

3. 大力构建算法开源与算力共享生态

制定相关政策，鼓励高校、研究机构和企业共同建设我国自主可控的开源平台，打破美国Github等开源社区霸王条款的束缚，引领全球开源技术，推动生成式AI和具身智能的自主发展。依托国家实验室、AI产教创新平台、头部企业等AI优势机构，为生成式AI与具身智能研究提供算力支撑。建立共享数据生态环境，降低数据获取成本，获取中文语料数据，将数据潜力最大化，推动我国生成式AI和具身智能的加速发展。

利益冲突声明

本文作者在此声明彼此之间不存在任何利益冲突或财务冲突。

Received date: July 10, 2025; **Revised date:** August 16, 2025

Corresponding author: Zheng Nanning is a professor from the Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, and a member of Chinese Academy of Engineering. His major research field is pattern recognition and intelligent systems. E-mail: nnzheng@xjtu.edu.cn

Funding project: Chinese Academy of Engineering project “Development Strategies and Countermeasures for Generative AI and Embodied Intelligence” (2024-XZ-14)

参考文献

[1] Turing A M. Computing machinery and intelligence [M]. Dordrecht: Springer Netherlands, 2007: 23–65.

[2] Wiener N. Cybernetics [J]. Scientific American, 1948, 179(5): 14–19.

[3] Brooks R A. Intelligence without representation [J]. Artificial Intelligence, 1991, 47(1–3): 139–159.

[4] Zador A, Escola S, Richards B, et al. Catalyzing next-generation artificial intelligence through NeuroAI [J]. Nature Communications, 2023, 14: 1597.

[5] Li F F, Krishna R. Searching for computer vision north stars [J]. Daedalus, 2022, 151(2): 85–99.

[6] Glenberg A M. Embodiment as a unifying perspective for psychology [J]. WIREs Cognitive Science, 2010, 1(4): 586–596.

[7] Ding N, Qin Y J, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models [J]. Nature Machine Intelligence, 2023, 5(3): 220–235.

[8] Zheng N N, Liu Z Y, Ren P J, et al. Hybrid-augmented intelligence: Collaboration and cognition [J]. Frontiers of Information Technology & Electronic Engineering, 2017, 18(2): 153–179.

[9] Durante Z, Huang Q Y, Wake N, et al. Agent AI: Surveying the horizons of multimodal interaction [EB/OL]. (2024-01-07)[2025-06-10]. arXiv: 2401.03568. https://arxiv.org/abs/2401.03568.

[10] Ha D, Schmidhuber J. World models [EB/OL]. (2018-03-27)[2025-06-10]. arXiv: 1803.10122. https://arxiv.org/abs/1803.10122.

[11] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798–1828.

[12] Pearl J. Causality [M]. New York: Cambridge University Press, 2009.

[13] OpenAI, Achiam J, Adler S, et al. GPT-4 technical report [EB/OL]. (2023-03-15)[2025-06-10]. arXiv: 2303.08774. https://arxiv.org/abs/2303.08774.

[14] Liu Y X, Zhang K, Li Y, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models [EB/OL]. (2024-02-27)[2025-06-10]. arXiv: 2402.17177. https://arxiv.org/abs/2402.17177.

[15] DeepSeek-AI, Liu A X, Feng B, et al. DeepSeek-V3 technical report [EB/OL]. (2024-12-27)[2025-06-10]. arXiv: 2412.19437. https://arxiv.org/abs/2412.19437.

[16] Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science [J]. Nature, 2023, 614(7947): 214–216.

[17] Team G, Anil R, Borgeaud S, et al. Gemini: A family of highly capable multimodal models [EB/OL]. (2023-12-19)[2025-06-10]. arXiv: 2312.11805. https://arxiv.org/abs/2312.11805.

[18] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436–444.

[19] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models [EB/OL]. (2020-01-23)[2025-06-10]. arXiv: 2001.08361. https://arxiv.org/abs/2001.08361.

[20] Heess N, Tb D, Sriram S, et al. Emergence of locomotion behaviours in rich environments [EB/OL]. (2017-07-08)[2025-06-10]. arXiv: 1707.02286. https://arxiv.org/abs/1707.02286.

[21] Liu Y, Chen W X, Bai Y J, et al. Aligning cyber space with physical world: A comprehensive survey on embodied AI [EB/OL]. (2024-07-09)[2025-06-10]. arXiv: 2407.06886. https://arxiv. org/abs/2407.06886.

[22] 中国信息通信研究院, 北京人形机器人创新中心有限公司. 具身智能发展报告(2024年) [R]. 北京: 中国信息通信研究院, 2024.
China Academy of Information and Communications Technology, Beijing Humanoid Robot Innovation Center Co., Ltd. Development report on embodied intelligence (2024) [R]. Beijing: China Academy of Information and Communications Technology, Beijing Humanoid Robot Innovation Center Co., Ltd., 2024.

[23] Zhao H S, Jiang L, Jia J Y, et al. Point transformer [R]. Montreal: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2022.

[24] 陶永, 万嘉昊, 王田苗, 等. 构建具身智能新范式: 人形机器人技术现状及发展趋势综述 [J/OL]. 机械工程学报, 2025: 1–27 [2025-05-08]. https://kns. cnki. net/KCMS/detail/detail. aspx? filename=JXXB202505060001&dbname=CJFD&dbcode=CJFQ.
Tao Y, Wan J H, Wang T M, et al. Establishing a new paradigm of embodied intelligence: A review of the current status and development trends in humanoid robot technology [J/OL]. Journal Of Mechanical Engineering, 2025: 1–27[2025-05-08]. https://kns. cnki. net/KCMS/detail/detail. aspx? filename=JXXB202505060001&dbname=CJFD&dbcode=CJFQ.

[25] 杨玉琪, 王梦云, 刘运卓, 等. 具身智能及其在自主无人系统的应用研究 [J]. 无人系统技术, 2024, 7(5): 99–110.
Yang Y Q, Wang M Y, Liu Y Z, et al. Embodied intelligence and its application in autonomous unmanned systems [J]. Unmanned Systems Technology, 2024, 7(5): 99–110.

[26] Ahn M, Brohan A, Brown N, et al. Do as I can, not as I say: Grounding language in robotic affordances [EB/OL]. (2022-04-05)[2025-06-10]. arXiv: 2204.01691. https://arxiv.org/abs/2204.01691.

[27] Brohan A, Brown N, Carbajal J, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control [EB/OL]. (2023-08-30)[2025-06-10]. arXiv: 2307.15818. https://arxiv.org/abs/2307.15818.

[28] Belkhale S, Ding T L, Xiao T, et al. RT-H: Action hierarchies using language [EB/OL]. (2024-03-06)[2025-06-10]. arXiv: 2403.01823. https://arxiv.org/abs/2403.01823.

[29] Wang H Q, Wang W G, Liang W, et al. Structured scene memory for vision-language navigation [R]. Nashville: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[30] Pan C B, Yaman B, Nesti T, et al. VLP: Vision language planning for autonomous driving [R]. Seattle: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[31] Kolve E, Mottaghi R, Han W, et al. AI2-THOR: An interactive 3D environment for visual AI [EB/OL]. (2017-12-14)[2025-06-10]. arXiv: 1712.05474. https://arxiv.org/abs/1712.05474.

[32] Savva M, Kadian A, Maksymets O, et al. Habitat: A platform for embodied AI research [R]. Seoul: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[33] Peng X B, Andrychowicz M, Zaremba W, et al. Sim-to-real transfer of robotic control with dynamics randomization [R]. Brisbane: 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018.

[34] Fang H S, Fang H J, Tang Z Y, et al. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot [EB/OL]. (2023-07-02)[2025-06-10]. arXiv: 2307.00595. https://arxiv. org/abs/2307.00595.

[35] Wan Z L, Ling Y G, Yi S L, et al. VinT-6D: A large-scale object-in-hand dataset from vision, touch and proprioception [EB/OL]. (2024-12-31)[2025-06-10]. arXiv: 2501.00510. https://arxiv. org/abs/2501.00510.

[36] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: A versatile and accurate monocular SLAM system [J]. IEEE Transactions on Robotics, 2015, 31(5): 1147–1163.

[37] Zhu X Y, Li J, Liu Y, et al. A survey on model compression for large Language Models [J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1556–1577.

[38] Wan Z W, Wang X, Liu C, et al. Efficient large language models: A survey [EB/OL]. (2024-05-20)[2025-06-10]. arXiv: 2312.03863. https://arxiv.org/abs/2312.03863.

[39] 郑南宁. 认知过程的信息处理和新型人工智能系统 [J]. 中国基础科学, 2000, 2(8): 9–18.
Zheng N N. Information processing for cognition process and new artificial intelligent systems [J]. China Basic Science, 2000, 2 (8): 9–18.

[40] LeCun Y. A path towards autonomous machine intelligence [J].

Open Review, 2022, 62(1): 1–62.

[41] 李强. 政府工作报告——2025年3月5日在第十四届全国人民代表大会第三次会议上 [J]. 工业信息安全, 2025 (2): 81–93.
Li Q. Report on the work of the government—Delivered at the third session of the 14th national people's congress of the people's republic of China on March 5, 2025 [J]. Industry Information Security, 2025 (2): 81–93.

[42] 中华人民共和国工业和信息化部.《"十四五"机器人产业发展规划》解读 [J]. 自动化博览, 2022, 39(3): 14–15.
Ministry of Industry and Information Technology of the People's Republic of China. Interpretation of *the development plan of robot industry in the* 14*th Five-Year Plan* [J]. Automation Panorama, 2022, 39(3): 14–15.

[43] DeepSeek-AI, Guo D Y, Yang D J, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning [EB/OL]. (2025-01-22)[2025-06-10]. arXiv: 2501.12948. https://arxiv.org/abs/2501.12948.

[44] Jaech A, Kalia A, Lerer A, et al. OpenAI o1 system card [EB/OL]. (2024-12-21)[2025-06-10]. arXiv: 2412.16720. https://arxiv. org/abs/2412.16720.

[45] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [R]. Las Vegas: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[46] Chang A, Dai A, Funkhouser T, et al. Matterport3D: Learning from RGB-D data in indoor environments [EB/OL]. (2017-09-18)[2025-06-10]. arXiv: 1709.06158. https://arxiv.org/abs/1709.06158.

[47] Ramakrishnan S K, Gokaslan A, Wijmans E, et al. Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI [EB/OL]. (2021-09-16)[2025-06-10]. arXiv: 2109.08238. https://arxiv.org/abs/2109.08238.