具身智能系统的“大脑”里都有什么？

蒋树强[1,2] 宋新航[2] 张思贤[1,2]

[1] 中国科学院大学 [2] 中国科学院计算技术研究所

具身智能机理与内涵

伴随着20世纪中叶计算机的发明，人工智能技术进入大家的视野，虽几经波折，仍得到长足发展，正在深刻影响社会发展与技术进步。如何利用计算技术实现模拟人、超越人的智能能力是人工智能领域的核心问题。近十年来，随着深度学习技术的突破和计算能力的提高，大模型技术得到快速发展，在提升图文工作效率和改善网络信息服务方面取得很大进步。然而，以深度神经网络为基础的大模型训练数据主要来自互联网，缺乏与真实物理世界的交互，没有可移动的传感器及其相对应的参照系，与真实世界的情境脱节，这种离身性导致其学到的知识有限”。人是综合的具身智能体，从生命起源的有机分子经过几十亿年进化而来，在大自然中生存，在动态场景中学习，与真实世界交互，能感知、会思考、知学习、善表达、具行为，是人工智能技术研究的重要比照对象。

具身智能作为人工智能的一个重要研究方向，近年来得到学术界和产业界的大量关注。具身智能是指通过物理智能本体与外界环境的互动来实现智能的理论与技术研究，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。涉身性最早是一个哲学概念，人类对世界的认知是通过与身体互动来实现的，身体的结构、感觉、运动系统对智能的形成和发展起着关键作用。在具身智能中，物理本体不仅是智能体的载体，也是信息获取的渠道与行为反馈的装置。具身智能的情境性涉及环境上下文，是指智能体在与环境交互过程中，离不开环境的时间、空间、对象、事件等各种上下文因素，也离不开智能体的位置、视角、任务和能力，这样才能更好理解情境并做出合适的决策与行动。主动性是具身智能区别于传统静态、离身人工智能的重要特性，智能体通过在环境中主动感知、主动探索、主动交互、主动行为和主动学习，以更好适应理解环境、提升自身能力、高效完成任务。具身智能体在完成任务的过程中无时无刻不与环境中的对象与智能体产生各种交互，以此获取感知变化，并通过自己的行为影响环境。由此可见，具身智能是智能、身体与环境三者之间存在紧密联系，相互关联与影响，以实现环境感知、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。

具身智能系统

具身智能系统作为一个综合性系统，包括本体形态、传感模块、计算模块和智能算法模块等部分，各个模块之间紧密关联并交叉融合，这涉及软硬件一体化、算法与程序实现的协同、数据与知识的结合、机械本体与计算体系结构的结合等诸多方面，形成了复杂而高效的综合智能体，如图1所示。

本体形态是具身智能系统的物理基础，既可以是地面上的机器人，也可以是空中或水下的无人机、无人艇、机器鱼，也可以是适应不同场景需求下灵活多样、不同形态的各种机械本体，它决定了系统与外界环境的交互方式。具身智能本体形态的设计需要充分考虑

DOI:10.11991/cccf.202508007

基金项目： 国家自然科学基金项目（62125207, U23B2012）；北京市自然科学基金项目（ L24202

0） **通信作者：** 蒋树强，E-mail: sqjiang@ict.ac.cn

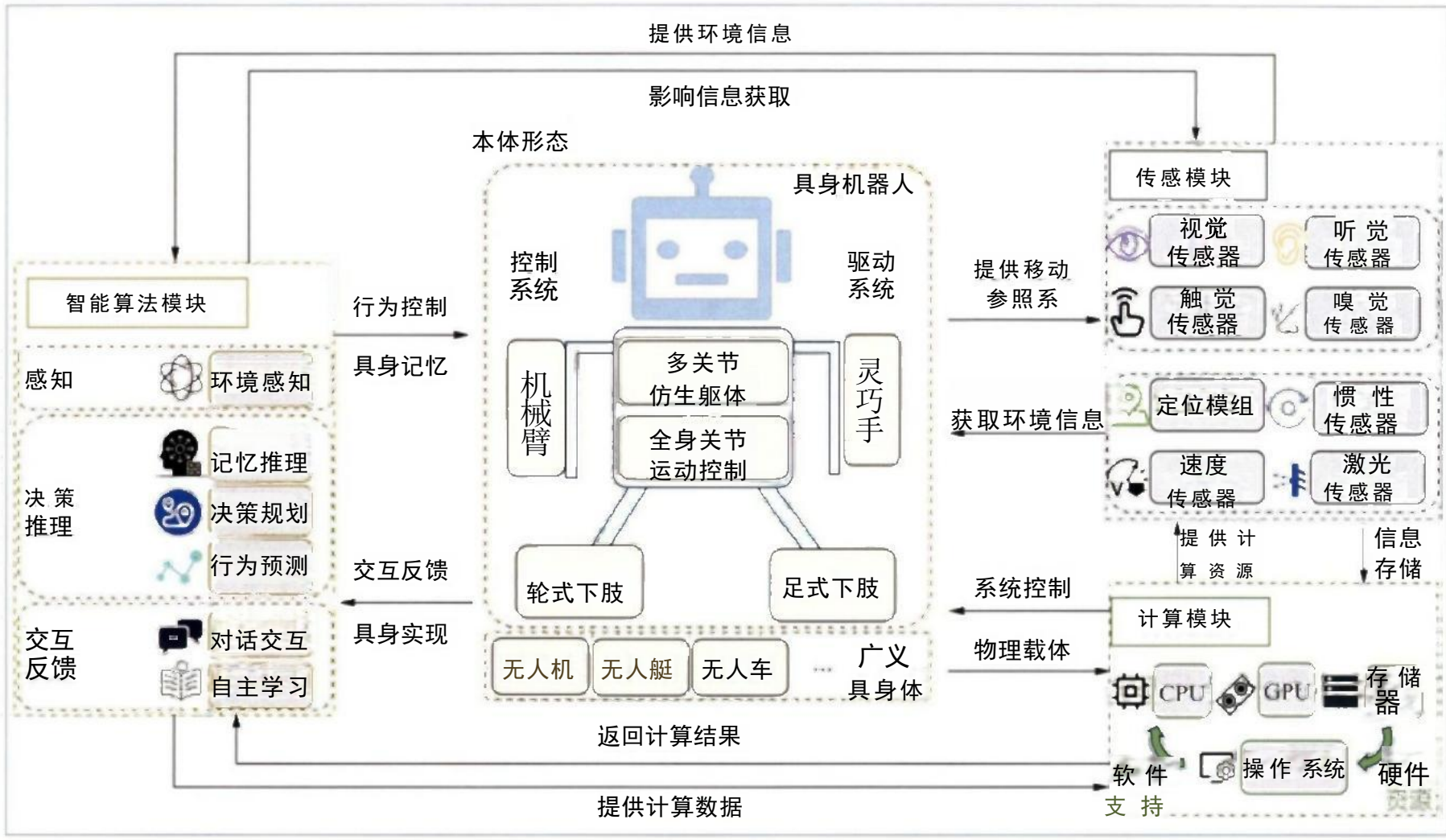

图 1 具身智能系统

系统的功能需求、环境适应性以及与其他模块的兼容性。传感模块是具身智能系统感知外界环境的重要途径，它能够收集各种信息，如视觉、听觉、触觉等，为系统的决策与行为提供基础的感知数据。传感模块的精度和可靠性直接影响到系统对环境的感知能力，进而影响到系统的决策和行动。计算模块是具身智能系统的能力底座，它负责对传感模块收集到的数据进行处理和分析，支持智能算法的运行，实现多任务的协调，并对具身本体执行系统控制，这需要具备强大的计算和负载能力。系统在运行过程中，可能会同时执行多个任务，如物体识别、抓取操作及路径规划等，这就要求计算系统能够承受高负载，以快速准确地处理大量的数据，稳定地运行各个任务。因此合理的软件算法和任务调度策略能充分发挥计算模块的性能，使具身智能系统更加高效和可靠。智能算法模块为具身智能系统提供了推理、决策和学习的能力，它能够根据传感模块收集到的信息，做出合理的决策，并通过学习不断优化自身的性能。当前以深度学习为代表的机器学习技术是智能算法模块中常用的解决方案，它通过大量的数据训练，采用有监督或自监督机器学习技术自动提取数据中的特征和规律，以实现环境理解、语言对话、行为规划、自主学习等能力。当然，智能算法涉及面广，除深度学习外，还会根据不同任务需求涉及强化学习、逻辑推理、进化算法等诸多技术，不同技术在具身智能系统中相互补充、协同工作，以更好适应各种复杂的应用场景。随着研究的深入，具身智能算法仍存在巨大的探索空间，未来发展前景广阔。

具身智能系统既要“具身”又要“智能”，因此多层面的软硬件结合是具身智能系统的典型特征，这需要系统既拥有可靠的机械本体与足够的计算支撑，又具备灵活的控制、决策与学习能力。要更好实现具身智能系统中的“智能”能力，需要更全面地探讨其“大脑”“小脑”和“本体”间的紧密耦合关系，更深入分析具身智能系统中的“大脑”能力，以更好模拟类人的智能、支持多个智能算法模块的协同，下一节将对具身智能系统的“大脑”能力及其实现进行深入剖析。

具身智能系统中的“大脑”能力与技术实现

具身智能系统“大脑”能力

在人类的具身活动中，“大脑”“小脑”与“本体”协同运作：大脑主导感知、记忆、学习与决策，小脑负责运

动协调与精细控制，而本体则提供与外部环境互动的物理基础。这种紧密协作机制为具身智能系统的设计提供了重要的生物学启发，尤其是人类大脑相关脑区在具身活动中，位于大脑半球中部，中央沟后方顶叶是人类感知的重要脑叶之一。顶叶感觉处理区接收并整合来自皮肤、关节和肌肉的触觉与感觉信号，角回负责处理空间关系与抽象符号，将三维环境的内在逻辑与数学符号转化为空间认知[2]。此外，位于大脑半球后部的枕叶，专注于视觉信息的深度解演，其中所包含的初级、次级和高级视皮层如同像素解析器，将光线分解为轮廓、色彩与运动轨迹。同时枕叶中的视觉处理区进一步识别物体概念与场景语义。

在学习与记忆层面，颞叶内的海马体是人脑的记忆中枢。海马体实现学习和记忆的生物学基础是神经的可塑性3。这种可塑性包含多种机制，长时程增强机制强化与情境信息相关的突触通路，而长期未强化的冗余连接则被长时程抑制修剪，使神经网络在感知刺激下不断重塑，这些机制让海马体将感官信息与情境细节编织成情景记忆。海马体相邻的杏仁核为记忆赋予情感权重并推动短期记忆向长期存储转化。同时记忆需要在学习后的离线阶段（如睡眠）通过“重放”活动来巩固46。通过人脑的学习与记忆的过程，人类就可以形成记忆并用学到的信息完成后续的决策任务7。

在决策与行动层面，近年来的研究表明，决策并非某一个脑区的“独角戏”，而是全脑范围内的共同任务。前额叶皮层作为参与决策层面的重要脑区8，可以评估信息价值并抑制冲动反应，同时调度工作记忆暂存多任务信息。海马体通过θ波节律实现和位于额叶的眼窝额皮质的同步，共同参与到决策当中9。由多个脑区结合产生的决策，会通过运动皮层等区域将决策转化为动作序列，最终完成行动过程。

为了使具身智能系统可以像人类一样完成具身活动，具身智能系统应包括感知、记忆与学习、决策三大核心功能模块，如图2(b)所示，它们彼此紧密协作，形成类似人脑的感知、记忆与学习、决策的闭环。

感知模块是系统获取外界信息的门户。它模拟人脑中枕叶和顶叶等区域的功能，负责采集和处理来自外部环境的多模态信息（如视觉、听觉、语言等），并结合上下文对信息进行解析。这一过程不仅为后续的记忆与决策提供输入，也通过实时感知支持短期反应和环境适应，是具身智能系统感知与理解世界的基础。

记忆与学习模块承担着对环境知识的抽象和积累，类似于人脑中海马体的作用，它通过对感知信息进行场景建模与经验总结，形成对环境规律和交互情境的内化知识。这些知识不仅包括对过去情境的存储，也支持通过主动学习不断优化，实现知识的迁移与泛化。

决策模块类似人脑额叶的功能，承担综合判断与

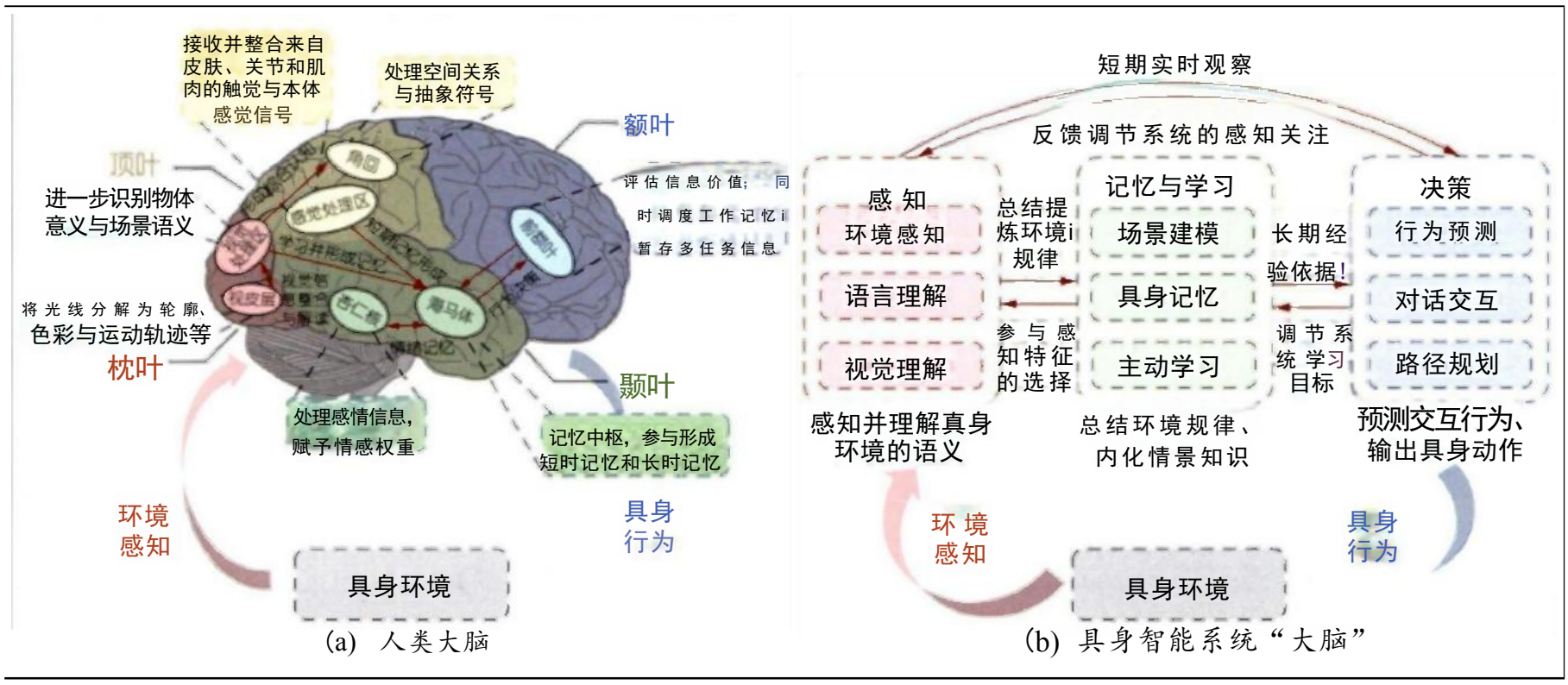

图2 人类大脑与具身系统“大脑”

行动输出。它将来自感知的实时信息与记忆中的长期知识进行融合，用于行为预测、任务规划、路径选择和语言交互等操作，从而驱动系统产生具体的身体行为。这一模块不仅是系统输出的“发动机”，也会根据 反馈调整未来的学习与感知重点。

这3个模块之间不仅存在信息传递关系，更构成了一个高度耦合、动态协同的整体认知结构。感知不仅接收信息输入，还是决策过程的依据；记忆不仅存储信息，还参与感知特征的选择与策略决策；决策不仅输出行为规划，还通过自主学习调节感知关注的重点。整个系统在具身活动过程中是一个不断循环迭代、自我调节的认知体。

此外，上述具身系统大脑中的智能功能与身体行为和环境之间同样紧密耦合，大脑和身体通过行为与感知的持续循环而动态连接在一起10]。感知不仅是对视觉和语言的理解，更是服务于环境与身体，即视觉理解要支持身体在环境中的避障，语言理解要引导身体在环境中的交互行为，身体行为也会反过来影响智能系统的认知状态。例如，身体位置和运动路径的变化会重新定义观察坐标系与感知范围，从而改变对场景的理解。这种耦合关系说明具身智能不只是大脑的产物，而是大脑、身体和环境整体协调的结果。这种耦合关系构成了具身智能系统独有的认知闭环，也是其区别于传统离身式的感知与推理系统的根本所在。

具身智能系统中的“大脑”技术实现

当前具身智能系统中多个关键技术正在快速发展与融合，支撑智能系统实现像人类一样的认知闭环，这些技术主要包括具身大模型、世界模型、具身记忆、行为预测与自主学习等。

近年来，具身大模型因其可以端到端地在复杂真实环境中通过感知信息来进行对应的决策行动而备受关注。与传统的人工智能系统不同，具身大模型集成了多种感官模态，如视觉、语言和音频等，使得智能体能够感知并与物理环境进行互动。近年来，随着大语言模型（large language model,LLM）与多模态感知系统 的发展，促使了一系列例如 RT-2!、OpenVLA[12)、π .1[3]、 Gemini Robotics 等新型具身大模型的开发。这些研 究聚焦于具身大模型的数据集、多模态对齐融合等方 向，使模型能够同时处理多模态输入，并输出与物理世界的交互动作。

此外，备受关注的还有世界模型研究领域。当前的研究核心是理解与建模变化中的世界，并对变化世界的下一时刻状态进行预测，主要侧重于两个方面：构建内部表征以理解世界并预测未来状态以模拟和指导决策。2022年，Yann LeCun[15]提出了一种联合嵌入预测架构，它由一个处理感官数据的感知模块和评估这些信息的认知模块组成。最近的研究也结合大语言模型、视觉模型等方法捕捉包括空间与时间知识，或是在模型内部嵌入类脑结构16，这使得模型可以根据先前 的经验学习预测未来事件。总体来看，世界模型的研 究仍处于起步阶段，不同研究从各自角度提出技术方 案，尚未形成统一范式，仍在不断演化与迭代之中，未 来还有很大的发展空间。

具身记忆领域模仿人脑的记忆机制，存储与环境交互过程中积累的经验、场景特征、任务状态等。它不仅记录“做过什么”，更记录“在哪种情境下做过什么”，支持决策模块在未来情境中快速匹配已有经验。当前具身记忆系统多通过显式的多层级图结构17-18或隐式的 Transformer架构19来实现，研究热点集中于记忆检索、跨模态统一表示以及记忆与策略学习的协同机制等方向，例如，OpenAI在引入了外部“长期记忆”机制[20,为大语言模型提供跨会话的记忆能力；Meta AI于2024年提出的V-JEPA 架构21则采用时空遮蔽策略，在抽象表示空间内预测视频中的被遮挡内容，有助于捕捉高级概念信息。

行为预测领域关注通过对历史及实时数据的分析，预判自身或其他实体的未来动作轨迹，从而解决物理环境中的动态不确定性。近年来，行为预测技术已从简单的轨迹推演扩展至复杂场景的动态建模。当前研究聚焦于如何通过高效的序列建模22、潜变量学习 1 2等技术，使智能体具备类人的预判能力，为智能体提供前瞻性决策依据。

自主学习领域通过智能体与环境的动态交互实现策略优化和能力演进，其核心包括强化学习、主动学习、终身学习、逻辑推理、进化算法等技术。其中，强化学习作为核心范式，驱动智能体在试错中优化决策策略，通过“状态—动作—奖励”的交互机制优化智能体的决策与行动策略，使其能自主学习复杂任务241。而主动学习也可以通过少量数据标注，引导视觉模型

达到更好的训练效果，提升智能体的感知能力[25]。

综上所述，尽管当前具身智能的研究在感知、记忆、学习、决策等方向上各有侧重，但这些能力并非孤立发展或简单叠加。多种技术通过信息流动、任务协同与反馈调节，共同构成了一个高度耦合、动态反馈的一体化智能系统。其中感知模块接收的多元环境信息，为记忆模型对当下环境理解与未来预测提供了实时输入；学习模型构建的内部分析与状态预测结果，又影响决策模块对未来的评估决策；而环境反馈又驱动着整个系统策略的持续优化，并持续提升感知精度与预测能力，形成一个紧密协作的闭环。

需要指出的是，尽管大语言模型在环境理解与任务决策中展现出强大的能力，但它并不能独立承担具身智能系统的全部“大脑”功能。大模型依赖于海量离线数据训练，具备丰富的语言与视觉先验知识。然而，这类模型本质上并非“具身”的，它们缺乏对实时环境的直接感知能力，也无法通过自主行为从环境中持续获取信息。具身智能系统面临的环境动态变化、不可预见、充满噪声，因此系统需要通过实际的传感器输入感知当下情境，并与世界持续互动。这些信息往往具有时序性、局部性和物理约束性，无法仅通过离线训练建模获得。因此，大模型可被视为“知识引擎”,能够在推理、理解、计划中发挥作用，但实现具身智能系统还必须依赖于对具身环境的感知、对情境的记忆积累以及对行动结果的反馈等技术协同，才能实现完整、闭环的智能能力。

具身智能技术发展趋势与展望

具身智能技术正经历从数据驱动到交互驱动、从被动感知到主动预测、从分层决策到一体化、从定向训练到自主进化的多维范式跃迁。传统模式依赖静态数据与预设规则，而未来通过多模态交互、环境动态反馈及端到端学习框架，使智能体在物理互动中自主“涌现”认知能力，并逐步建立环境因果模型与长期预测机制。决策架构突破线性时序限制，融合强化学习与量子启发算法实现复杂约束下的实时规划；学习方式则向元学习与跨域进化演进，结合数字孪生与伦理约束框架，推动智能体从环境适应者转向共塑者。这四大技术趋势共同勾勒出具身智能从“模拟智能”向“实体智能”跨越的发展图景，标志着具身智能正从实验室走 向真实物理世界的深刻转型。

智能涌现模式从数据驱动到交互驱动

具身智能技术正经历从数据驱动向交互驱动的范 式跃迁。传统数据驱动模式依赖海量标注数据训练模 型，但存在数据瓶颈与场景泛化难题。随着多模态大 模型与仿真环境的融合，具身智能体开始通过从环境 感知到接收环境实时反馈的闭环实现交互驱动学习。 例如，人形机器人通过3D视觉和触觉感知与环境动态 交互，在仿真环境中试错优化运动策略，再通过Sim2Real技术迁移至物理世界，显著提升复杂场景适 应能力。交互驱动的核心在于让智能体在物理互动中 “涌现”出认知能力。未来，具身智能将进一步融合神 经形态计算与世界模型，使机器人能预演动作后果，实 现从被动执行到主动探索的进化，推动工业制造、医疗 康复、家庭服务等领域向自主决策与零样本学习迈 进。这一转变不仅重构了智能“涌现”逻辑，更标志着认知能力具身化的深刻变革。

具身理解机制从实时感知到预测未来

具身智能技术正推动从被动感知到主动预测的理 解机制突破。传统感知依赖预设规则，而具身交互驱 动智能体通过视觉、触觉等多模态传感器主动探索环 境，形成动态状态理解。进一步，具身智能体基于实时 反馈与历史经验，逐步建立环境和行为因果模型，实现 从即时感知到短期预测的跃迁。例如，通过强化学习 优化运动策略，具身智能体可预判物体滑落风险并提 前调整抓取力度。未来，随着世界模型与因果推理的 融合，具身智能体将具备长期预测能力，在复杂场景中 自主规划多步决策。这一过程不仅突破了感知局限， 更能推动智能体从环境适应者向环境塑造者演进，为 开放场景下的自主作业奠定基础。

具身决策与规划从分层到一体化

具身智能正经历从分层决策向一体化决策的范式 转变。传统分层架构通过感知、规划、执行等分离模块 实现可控性，但存在实时性不足与数据对齐难题。 一 体化决策通过端到端深度学习框架，将多模态传感器 数据（如视觉、触觉、本体感知）直接映射至决策和动作

空间，利用端到端网络架构实现跨模态特征融合与长程依赖建模。此外，强化学习与模仿学习的混合架构可平衡探索效率与动作精度，在复杂约束场景下实现快速响应。这种范式转变突破了分层架构的线性时序瓶颈，通过共享表征学习可提升数据利用率，但仍面临计算复杂度指数级增长与可解释性下降的挑战。未来，可融合新原理启发式算法，提升复杂约束下的并行决策效率和边界上限，结合神经符号系统构建可解释的决策策略模型，在保证实时性的同时提升长程规划能力。

具身学习从定向训练到自主进化

具身学习正经历从定向训练到自主进化的范式跃迁 。传统定向训练依赖人工设计的任务场景与奖励函数，通过监督学习或强化学习实现技能习得，但存在环境适应性弱、迁移成本高的瓶颈。当前研究聚焦于构建具备元学习能力的自主进化框架，通过神经符号系统整合逻辑规则与深度表征，使具身智能体能在动态环境中自主构建任务序列，结合因果推理实现跨域知识迁移与进化。同时，具身智能体通过虚拟—现实交互 产生的海量数据流，通过神经辐射场（neural radiance fields, NeRF）或三维高斯模型与大语言模型的深度耦 合，构建物理世界数字孪生体，实现虚拟—现实语义对 齐与跨模态因果推理。未来，具身学习将深化多模态 交互与跨域协同进化，推动智能体从环境响应者进化 为环境共塑者，再基于可解释 AI 的进化约束框架通过 形式化语言构建安全沙盒，确保自主决策过程既符合 物理世界运行规律，又满足真实应用价值要求，最终实 现人机物三元空间的智能融合与协同进化。

展望未来，具身智能技术的四大发展趋势将形成协同效应，推动智能系统实现从“环境适应”到“环境交互与塑造”的质变。交互驱动的认知构建将突破数据瓶颈，预测性理解机制将拓展决策维度，融合强化学习的启发式决策算法加速复杂空间求解，并通过具身元学习框架实现自主进化。这些变革不仅将重塑工业制造、医疗康复等传统领域，更将在深空探测、深海作业等极端场景中开辟全新应用可能。未来，不仅具身“大脑”技术将会跨越式进化，当具身智能体能够自主进化形态、预判环境变化、群体协同完成复杂任务时，具身智能将真正成为连接数字世界与物理世界的“新生命形态”，开启人工智能与人类社会协同进化的新纪元。需要指出的是，人类大脑本身仍然极其复杂，其运行机制尚未被完全揭示，因此模拟与重构这一过程的具身智能技术仍处于持续探索阶段，不同方向的技术尝试不断涌现，整体发展潜力巨大，未来仍有广阔的深入研究空间。

蒋树强

CCF 杰出会员、监事、多媒体技术专业委员会副主任。中国科学院大学特聘教授。主要研究方向为多媒体技术与具身智能。sqjiang@ict.ac.cn

宋新航

CCF 专业会员。中国科学院计算技术研究所副研究员。主要研究方向为具身智能、视觉导航。xinhang.song@ict.ac.cn

张思贤

中国科学院计算技术研究所助理研究员。主要研究方向为具身智能、视觉导航 。sixian.zhang@vipl.ict.ac.cn

参考文献

[1]Jeff Hawkins. *A Thousand Brains:A New Theory of* ***Intelligence***[M].New York:Basic Books,2021.

[2]M.A.Goodale,A.D.Milner,L.S.Jakobson,et al.A Neurological Dissociation Between Perceiving Objects and Grasping Them[J].*Nature*,1991,349(6305):154-156.

[3]Yuhang Song,Beren Millidge,Tommaso Salvatori,et al. Inferring Neural Activity Before Plasticity as a Foundation for Learning Beyond Backpropagation[J]. *Nature Neuroscience,*2024,27(2):348-358.

[4]John Lisman,A.D.Redish.Prediction,Sequences and the Hippocampus[J]. *Philosophical Transactions of the Royal Society of London Series B,Biological Sciences*,2009, 364(1521):1193-1201.

[5]Wenbo Tang,Justin D Shin,Shantanu P Jadhav.Geometric Transformation of Cognitive Maps for Generalization Across Hippocampal-Prefrontal Circuits[J]. *Cell Reports,* 2023,42(3):112246.

[6]Wangjing Yu,Asieh Zadbood,Avi J H Chanales,et al.

Repetition Dynamically and Rapidly Increases Cortical,But Not Hippocampal,Offline Reactivation[J]. *Proceedings of the National Academy of Sciences of the United States of* America,2024,121(40):e2405929121.

[7]Aleksander P.F.Domanski,Michal T.Kucewicz,Eleonora Russo,et al.Distinct Hippocampal-Prefrontal Neural Assemblies Coordinate Memory Encoding,Maintenance, And Recall[J].Current Biology,2023,33(7):1220-1236.e4.

[8]Yael Niv.Learning Task-State Representations[J].Nature *Neuroscience,*2019,22(10):1544-1553.

[9]Thomas W.Elston,Joni D.Wallis.Context-Dependent Decision-Making in the Primate Hippocampal-Prefrontal Circuit[J].*Nature Neuroscience*,2025,28(2):374-382.

[10]Ogi Ogas,Sai Gaddam.*Journey of the Mind:How Thinking Emerged from Chaos[M]*.New York:WW Norton & Company,2022.

[11]Anthony Brohan,Noah Brown,Justice Carbajal,et al.Rt-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control[EB/OL].(2023-07-28)[2025-07-18]. https://arxiv.org/abs/2307.15818.

[12]Moo Jin Kim,Karl Pertsch,Siddharth Karamcheti,et al. Openvla:An Open-Source Vision-Language-Action Model [EB/OL].(2024-06-13)[2025-07-18].https://arxiv.org/abs/ 2406.09246.

[13]Kevin Black,Noah Brown,Danny Driess,et al.π_0 :A Vision-Language-Action Flow Model for General Robot Control[EB/OL].(2024-10-31)[2025-07-18].https://arxiv. org/abs/2410.24164.

[14]Gemini Robotics Team,Saminda Abeyruwan,Joshua Ainslie,et al.Gemini Robotics:Bringing AI into the Physical World[EB/OL]. (2024-10-31)[2025-07-18]. https://arxiv.org/abs/2503.20020.

[15]LeCun Yann.A Path Towards Autonomous Machine Intelligence Version 0.9.2[EB/OL].(2022-06-27)[2025-07-18].https://openreview.net/pdf?id=BZ5alr-kVsf.

[16]Yuxiao Li,E.J.Michaud,D.D.Baek,et al.The Geometry of Concepts:Sparse Autoencoder Feature Structure[J]. *Entropy*,2025,27(4):344.

[17]Sixian Zhang,Xinhang Song,Yubing Bai,et al.Hierarchical Object-To-Zone Graph for Object Navigation[C]//2021 *IEEE/CVF International Conference on Computer Vision.* Piscataway:IEEE,2021:15110-15120.

[18]Sixian Zhang,Xinhang Song,XinyaoYu,et al.HOZ: Versatile Hierarchical Object-To-Zone Graph for Object Navigation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,2025,47(7):5958-5975.

[19]Sixian Zhang,XinyaoYu,Xinhang Song,et al.Imagine Before Go:Self-Supervised Generative Map for Object Goal Navigation[C]//*2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Piscataway:IEEE,2024: 16414-16425.

[20]OpenAI,Josh Achiam,Steven Adler,et al.GPT-4 Technical Report[EB/OL].(2023-08-27)[2025-07-18].https://arxiv. org/abs/2303.08774.

[21]Adrien Bardes,Quentin Garrido,Jean Ponce,et al. Revisiting Feature Prediction for Learning Visual Representations from Video[EB/OL]. (2024-08-27) [2025-07-18].https://arxiv.org/abs/2404.08471.

[22]Lili Chen,Kevin Lu,Aravind Rajeswaran,et al.Decision Transformer: Reinforcement Learning via Sequence Modeling[EB/OL].(2021-06-02)[2025-07-18].https://arx iv.org/abs/2106.01345.

[23]Danijar Hafner,Timothy Lillicrap,Jimmy Ba,et al.Dream to Control:Learning Behaviors by Latent Imagination [EB/OL].(2019-12-03)[2025-07-18].https://arxiv.org/abs/ 1912.01603.

[24]Alisson Azzolini,Junjie Bai,Hannah Brandon,et al. Cosmos-Reason1:From Physical Common Sense to Embodied Reasoning[EB/OL]. (2025-03-18)[2025-07-18].https://arxiv.org/abs/2503.15558.

[25]Ming Xie,Yuxi Li,Yabiao Wang,et al.Learning Distinctive Margin Toward Active Domain Adaptation[C]//2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*Piscataway:IEEE,2022:7983-7992.

What Constitutes the "Brain"of an Embodied Intelligence System?

Shuqiang Jiang[2], Xinhang Song[1 2], Sixian Zhang[2]

1.University of Chinese Academy of Sciences

2.Institute of Computing Technology,Chinese Academy of Sciences

Abstract: Embodied intelligence refers to the intelligence emerging from the close coupling between an agent's physical body and its environment.By highlighting intelligence as shaped by bodily interactions,it inherently possesses characteristics of embodiment, situatedness,proactivity,and interactivity,and it is considered essential for effectively applying artificial intelligence within the physical world.Regarding the silicon-based "brain"of embodied intelligence systems,it is expected not only to control the physical body but also to perceive the environment,memorize contextual information,and plan actions.To enhance these capabilities,the "brain"of an embodied intelligence system should integrate visual observations,historical contexts,and prior knowledge.Additionally,it should be capable of envisioning future scenarios and adapting to its environment.These capabilities closely align with current technologies such as embodied foundation models,world models,In this paper,we first outline the definition and characteristics of embodied intelligence,then analyze the framework of embodied intelligence systems and review functional modules along with their interrelations.Furthermore,we discuss related technologies,including embodied foundation models,world models,embodied memory,action planning,and humanoid learning.