**多模态智能领域研究：具身智能中视觉-语言-动作协同及其应用**[[1]](#footnote-0)

摘 要：本文聚焦自然语言处理（多模态学习）、智能机器人（具身智能）及人机交互（多模态融合与脑机接口）领域，重点探讨具身智能中视觉 - 语言 - 动作协同在虚拟现实（VR）和医疗影像分析中的应用。通过梳理核心研究方向、强势院校及代表人物，结合技术进展与市场需求，分析该领域就业趋势与挑战，为相关研究与实践提供参考。

关键词：具身智能；视觉 - 语言 - 动作协同；虚拟现实；医疗影像分析；人机交互

**一、核心研究方向解析**

**（一）自然语言处理（多模态学习）**

多模态学习突破单一文本处理局限，整合文本、图像、音频、视频等多源数据，实现更精准的语义理解与生成。核心方向包括：

1． 跨模态语义对齐：通过深度学习模型（如 Transformer）构建多模态统一表征空间，解决文本与视觉 / 听觉信号的语义映射问题。例如，结合医学影像与电子病历文本，辅助肿瘤诊断时精准关联病灶特征与临床描述。

2． 多模态生成任务：基于大规模预训练模型（如 DALL-E、GPT-4）实现图文互生成、视频描述生成等，推动创意设计、教育辅助等场景落地。

3． 情感与意图融合分析：融合语音语调、面部表情与文本情感，提升人机对话系统的情感理解能力，如智能客服对用户情绪的实时响应。

**（二）智能机器人（具身智能与类人化设计）**

具身智能强调物理实体通过感知 - 决策 - 行动闭环与环境交互，核心研究方向包括：

1． 多模态具身感知：整合视觉（如双目相机）、触觉（压力传感器）、听觉（麦克风阵列）等传感器数据，构建环境语义地图。例如，人形机器人通过视觉识别物体位置，结合触觉反馈调整抓取力度，实现对玻璃器皿的安全操作。

2． 类人化动作规划与控制：模仿人类运动学原理，设计高自由度机械结构（如 55 自由度人形机器人 STAR1），结合强化学习优化步态算法，实现复杂地形行走（如 3.6m/s 奔跑速度）与精细操作（如使用筷子夹取食物）。

3． 视觉 - 语言 - 动作协同：通过分层决策模型（如 OpenAI 的 Figure 01 三层架构）或端到端模型（如 Google RT-2），将语言指令（如 “将药箱放到床头柜”）转化为视觉定位、路径规划与机械臂动作，解决具身智能的跨模态指令理解与执行问题。

**（三）人机交互（多模态融合与脑机接口）**

1． 多模态融合交互技术：整合语音、手势、眼动等交互方式，构建自然化人机界面。例如，智能车载系统中，用户通过语音指令 “导航至医院”，结合手势滑动选择路线，系统通过视觉识别驾驶员注意力状态调整交互优先级。

2． 脑机接口（BCI）交互：通过非侵入式（EEG）或半侵入式（柔性电极）技术采集脑电信号，实现神经信号到设备控制的直接映射。复旦大学研发的 “三合一” 颅骨植入式设备，术后 24 小时患者即可通过脑电信号控制下肢肌群，准确率达 92.7%，为瘫痪患者提供运动康复新路径。

**二、 强势院校与研究机构及代表人物**

**（一）国内院校与机构**

1. 清华大学

清华大学在具身智能领域的研究具有显著优势，其交叉信息研究院（IIIS）和智能产业研究院（AIR）是该领域的核心力量。交叉信息研究院下设多个实验室，包括陈建宇教授领导的智能系统与机器人实验室（ISR Lab），专注于人形机器人、强化学习与多模态控制算法，其团队开发的“DWL”算法实现了人形机器人在复杂地形中的稳定运动，并推动具身大模型与工业场景结合。许华哲教授组建的具身智能实验室（TEA Lab）则致力于机器人策略泛化性研究，提出的“3D扩散策略”算法显著提升了机器人操作的灵活性与泛化能力。此外，高阳教授领导的具身视觉与机器人实验室（EVAR Lab）聚焦具身智能与计算机视觉的融合，推动端到端强化学习框架的落地应用。

1. 浙江大学

浙江大学在具身智能基础理论与技术路线图方面表现突出，2025年发布的《具身智能白皮书》系统梳理了多模态感知、具身大模型等关键方向。杨易教授团队通过具身大模型推动工业机器人从单一任务执行向复杂环境自主决策升级，其研究成果在智能制造领域具有重要影响力。

1. 复旦大学

复旦大学于2025年3月成立的可信具身智能研究院，整合了计算机视觉、机器人学、科技伦理等多学科资源，重点布局五大方向：基础模型、数据引擎、具身交互、本体研制与可信机制。该研究院与多家企业共建联合实验室，推动医疗康养与制造业服务机器人落地。张嘉漪团队在脑机接口领域取得突破，开发纳米线人工光感受器修复视觉功能，宋恩名团队则攻克柔性侵入式电极材料难题，显著提升神经信号采集效率。

1. 上海交通大学

上海交通大学在医疗机器人领域具有显著优势，卢策吾教授团队开发的具身智能基础模型在手术导航与仿生机器人设计中实现突破，杨明教授主导的微创手术机器人项目已进入临床验证阶段。

**（二）国际院校与机构**

1. 卡内基梅隆大学（CMU）

CMU机器人研究所（RI）以传统机器人学与新兴学习方法的结合见长。侯一凡教授的研究聚焦于接触模式鲁棒性控制，其提出的量化评估方法显著提升了多接触操作的稳定性；杨硕教授则致力于足式机器人长距离定位与优化控制，将航天领域的数学理论应用于机器人轨迹规划。此外，CMU的人机交互研究所（HCII）在多模态交互系统（如眼动追踪与手势识别融合技术）上处于国际领先地位。

1. 斯坦福大学

斯坦福大学以具身智能学习算法为核心，李飞飞团队开发的视觉-语言预训练框架（如VIMA系统）大幅增强机器人对复杂指令的理解能力，推动具身智能向通用化发展。吴佳俊团队与许华哲教授合作，在无仿真物理真机强化学习领域取得突破。

1. 加州大学伯克利分校（UC Berkeley）

UC Berkeley的BAIR实验室在机器人动力学与多模态交互领域具有深厚积累，归国学者陈建宇、许华哲等人将伯克利的学术成果引入国内，创立星动纪元、星海图等企业，推动具身智能技术产业化。

**三、具身智能中视觉 - 语言 - 动作协同的应用方向**

**（一）虚拟现实（VR）场景**

1. 沉浸式交互体验：k用户通过自然语言指令（如 “举起右手”）结合视觉手势，控制虚拟化身完成复杂动作，如虚拟手术训练中精准操作器械。具身智能机器人 STAR1 的类人化动作库，为 VR 角色提供逼真运动模型，提升交互沉浸感。

2. 虚拟环境自主探索：机器人通过视觉识别虚拟场景物体（如门、楼梯），结合语言指令 “探索房间”，自主规划路径并执行开门、上下楼梯等动作，应用于虚拟展厅导览、灾害应急模拟训练。

**（二）医疗影像分析与临床应用**

1. 多模态诊断协同：整合 CT/MRI 影像视觉特征、病理报告文本与患者病史，通过视觉 - 语言模型自动生成诊断建议。例如，中南大学智能诊室机器人系统，结合视（影像）、触（病灶形态）、听（语音）信息，实现类人水平的体格检查与跨模态病征识别。

2. 手术导航与康复辅助：具身智能机器人依据术前影像规划手术路径，医生通过语言指令调整操作细节，视觉系统实时反馈组织特征，实现精准病灶切除（如神经外科手术）。复旦大学 “颈七神经交叉移位术 + 脑机接口” 联合方案，帮助偏瘫患者通过视觉想象控制机械臂，恢复肢体运动功能。

**四、就业趋势与市场分析**

**（一）就业场景与岗位需求**

1. 技术研发类：算法工程师（多模态融合、具身控制）、机器人系统工程师（机械设计与运动控制）、脑机接口研发工程师（神经信号处理），主要集中于科技企业（如 OpenAI、百度）、机器人公司（波士顿动力、大疆）及医疗科技企业（联影医疗、品驰医疗）。

2. 应用落地类：智能交互设计师（多模态界面设计）、医疗 AI 产品经理（辅助诊断系统推广）、康复治疗技术支持（脑机接口设备调试），就业于医疗、教育、智能家居等领域。

**（二）薪资待遇与竞争态势**

1. 薪资水平：算法工程师年薪普遍 20-50 万元，资深者可达 80 万元以上；机器人研发工程师初薪 15-30 万元，具备脑机接口经验者溢价 30%；智能交互设计师年薪 12-30 万元，跨模态设计能力突出者薪资上浮 20%。

2. 就业难度：核心岗位竞争激烈，需掌握 Python/TensorFlow/PyTorch 等工具，具备跨学科知识（计算机视觉 + 医学影像处理、自然语言处理 + 机器人学）。具备顶会论文（如 CVPR、NeurIPS）或知名项目（如人形机器人研发）经验者更受青睐。

**（三）市场规模与发展前景**

据赛迪顾问数据，2025 年全球具身智能市场规模预计达 248 亿美元，中国市场年增速超 30%，医疗康复、智能制造、居家服务为主要增长点。政策层面，国务院将具身智能纳入未来产业培育方向，北京、上海等地规划 2030 年实现核心技术自主可控，推动万套级机器人规模化应用。

**五、结论**

具身智能中视觉 - 语言 - 动作协同技术，通过融合多模态感知、类人化决策与自然交互，在虚拟现实与医疗影像分析中展现出显著应用价值。未来需突破跨模态融合效率、实时决策鲁棒性及伦理合规性等挑战，依托清华大学、浙江大学、复旦大学等高校的产学研协同，加速技术落地。就业市场对复合型人才需求旺盛，具备多学科背景与实践经验者将占据竞争优势，推动该领域从理论研究向普惠应用深度发展。

**参考文献**

1. 上海科技大学 .(2024, 12 月 24 日) .从具身感知到具身认知 —— 多模态大模型驱动的具身智能演进.上海科技大学官网. https://www.shanghaitech.edu.cn/2024/1223/c9a1104961/page.htm
2. CTO.COM. (2024, 12 月 6 日). 自然语言处理 (NLP): 开启人机交互新篇章. https://www.51cto.com/article/803404.html
3. 复旦大学周葆华,吴雨晴. (2024, 3 月). 超越单一模态:多模态计算传播研究的进展与前瞻.复旦大学传播与国家治理研究中心. https://cics.fudan.edu.cn/25/13/c40201a664851/page.htm
4. 北石化人工智能研究院. (2024, 12 月 13 日). 多模态大模型的开放环境适应性研究. https://ai.bipt.edu.cn/xwdt/hdxx/09d90ac9365c45ce9cdc0c628abc6734.html
5. 飞书 Z 研究. (2024, 1 月 22 日). 视觉 - 语言理解模型的当前技术边界与未来应用想象. https://www.feishu.cn/wiki/OurqweG2AiDnaxkyxgMcJM26nUg
6. 北京市人民政府. (2025, 2 月 28 日). 北京具身智能科技创新与产业培育行动计划（2025—2027 年）. 首都之窗.https://www.beijing.gov.cn/ywdt/gzdt/202503/t20250301\_4022474.html
7. 中国人工智能学会. (2025, 3 月). 中国人工智能学会具身智能白皮书.第二届 “中国具身智能大会” 发布.
8. 赛迪顾问. (2025, 1 月). 2025 全球具身智能市场发展白皮书.
9. 清华大学交叉信息研究院. (2025, 3 月). 具身智能通用平台 “慧思开物” 技术白皮书.
10. 中国政府网. (2025, 3 月 6 日). 具身智能如何走向未来？政策解读.https://www.gov.cn/zhengce/202503/content\_7010979.htm
11. 中关村论坛组委会. (2025, 4 月 1 日). 2025 中关村论坛年会人工智能主题日报告.
12. 复旦大学可信具身智能研究院. (2025, 3 月). 医疗具身智能白皮书:脑机接口与康复机器人应用研究.
13. 新华网. (2025, 1 月 3 日). 具身智能：人工智能的下一个浪潮.https://www.xinhuanet.com/milpro/20250103/5a2126336fe54c7590c9fd47e9a53454/c.html

1. [↑](#footnote-ref-0)