基于多模态大模型的具身智能体研究进展与展望

赵博涛，亢祖衡，瞿晓阳，彭俊清，张旭龙，王健宗
平安科技（深圳）有限公司，广东 深圳 518063

摘要
具身智能体指能够根据指令完成某种或多种任务并且具备与物理环境交互能力的智能实体。其在服务机器人、智能教育、辅助医疗等领域具有巨大的潜在应用，是实现通用机器人的重要途径之一。随着多模态大模型的发展，具身智能体具备了更强的语言理解、推理判断和环境感知能力，极大地推动了该领域的发展。近年来，具身智能体领域涌现出许多优秀的研究工作，但缺乏系统的调查评述。为了帮助研究者更全面地了解这一领域，对具身智能体的研究进行了深入调研与展望。首先，介绍了多模态大模型，其次回顾了常用数据集和用于构建具身智能体的物理载体。然后，回顾了具身智能体的3个关键研究方向：具身大模型、高级任务规划和低级动作控制。最后，总结了具身智能体领域面临的挑战和存在的局限性，并展望了未来的发展方向。该综述为研究者提供了有价值的参考，旨在促进具身智能体领域的进一步发展与创新。

关键词
具身智能体；多模态大模型；机器人；视觉语言模型；具身智能

中图分类号：TP39 文献标志码：A doi：10.11959/j.issn.2096–0271.2025035

Review and emerging trends of embodied agent based on multimodal large language models

ZHAO Botao, KANG Zuheng, QU Xiaoyang, PENG Junqing, ZHANG Xulong, WANG Jianzong
Ping An Technology (Shenzhen) Co. , Ltd. , Shenzhen 518063, China

Abstract
Embodied agents refer to intelligent entities capable of completing one or multiple tasks based on instructions and possessing the ability to interact with the physical environment. These agents have immense potential applications across various fields, such as service robotics, intelligent education, and assistive healthcare, and represent a crucial pathway toward realizing general-purpose robots. With the advancement of multimodal large language models, embodied agents possess enhanced abilities in natural language understanding, reasoning, and environmental perception, significantly accelerating progress in this domain. Although many outstanding works have emerged in recent years, the field still lacks comprehensive surveys and targeted evaluations. To help researchers quickly and thoroughly know the developments in this area, in-depth review and analysis were conducted. Multimodal large language models were introducted, followed by datasets and a review of the physical carriers used for

constructing embodied intelligent agents. Then, three key research directions are analyzed, including embodied large models, high-level task planning, and low-level action control. Finally, the challenges and limitations of embodied agents were summarized and potential future directions were explored. This review serves as a foundational reference for the research community and fosters further development and innovation in the field.

Key words
embodied agent, multimodal large language model, robot, vision-language model, embodied intelligence

0 引言

具身智能体指能够根据指令完成某种或多种任务并且具备与物理环境交互能力的智能实体。具身智能体作为一个跨人工智能、机械自动化、认知心理学等多个领域的重要研究方向，旨在赋予智能体在物理环境中感知、交互和行动的能力[1-2]。相较于传统的人工智能系统，具身智能体通过具象化的形态（如机器人、无人机等实体）直接与环境交互，实现更加自然和动态的任务执行方式。这种智能体不仅仅依赖于算法的计算能力，还通过自身的传感器、执行器以及物理形态与外部环境进行交互。也就是说，具身智能体的核心理念是，智能的形成不仅依赖于强大算力支撑下的大脑（即认知过程），还取决于多种传感器执行器支撑下的身体与环境的交互作用。从广义上来看，具身智能体可以被定义为具备物理载体（形态），能够通过感知-决策-执行的闭环系统完成与环境交互的智能体。其研究范畴涵盖自主机器人、多模态感知、强化学习、行为规划和人机交互等多个方向。随着研究的不断深入，具身智能体在服务机器人、智能教育和医疗辅助等领域展现出巨大的应用潜力[3-4]。

具身智能体的发展历史可以追溯到人工智能的早期探索。1950年，Turing[5]提出了“机器是否能思考”的问题，并预见了人工智能的终极形态：智能体应具备与环境交互感知、自主规划、决策、行动和执行的能力。但遗憾的是，在早期发展阶段，具身智能体的研究并不突出，因为当时的研究重点主要集中在人工智能在虚拟世界的计算推理能力上。1991年，Brooks[6]提出了“行为主义智能”，强调智能行为可以直接从自主机器与其环境的简单物理交互中直接产生，而这种交互不依赖于预先设定的复杂算法。Smith等[7]在2005年提出的“具身假说”从认知科学的角度对如何实现具身体智能体提出了6点要求，包括多模型、增量学习、具身性、主动探索、社交、学习一门语言。2018年，Ay等[8]从测度论的角度基于马尔可夫假设对具身智能体进行了抽象化的数学定义。同年，Laskey[9]在Stapp理论的基础上提出了一套理论，描述具身智能体如何在物理世界中采取行动以实现预期目标。实践层面，在具身智能体的早期阶段，研究人员主要集中在导航和定向任务，其中用到了很多计算机视觉的技术[10-11]。同时，还有部分研究人员从神经动力学的角度构建具身智能体，并取得了一定的进展[12-13]。然而，上述的方法在泛化性、通用性上仍然存在一定的局限性，离真正落地还存在较大的差距。近期，多模态大模型（multimodal large language model，MLLM）的出现为具身智能体的发展带来了一些新的突破，离实现真正意义上的具身智能体又更进一步。

多模态大模型是一种能够同时处理多种数据类型（如图像、文本、语音等）的先进人工智能模型。通过结合不同模态的数据，MLLM实现了更深层次的语义理解和跨模态推理，极大地提升了模型在视觉理解、语言处理以及多模态任务中的综合表现。MLLM能够将不同模态（如图像和文本）中的信息进行统一表示和融合，使模型能够从多种数据源中获取互补的信息，从而增强感知和理解能力。这类模型还可以在多模态数据之间建立关联，通过一种模态数据（如图像）生成或预测另一种模态数据（如文本），实现跨模态的理解与推理，在视觉问答、图像生成、语音识别、对话系统、内容推荐等任务中表现突出[14-15]。MLLM的典型代表包括CLIP[16]、DALL-E[17]以及视觉模型与语言模型相结合的视觉语言模型（vision-language model，VLM），如OpenAI的GPT-4o、谷歌的Gemini 1.5、国内推出的Qwen-VL系列[18]以及开源的LLaVA系列[19]等。

现有的多模态大模型展现出强大的环境感知、指令理解以及推理判断能力[15]，为具身智能体的发展注入了新的动力，并为实现真正的通用机器人带来了曙光。通过将MLLM的各项能力具象化到物理世界，具身智能体能够获得更强的任务执行能力。为了让具身智能体完成现实世界中复杂的任务，具身智能体被赋予了3个方面的能力。一是环境感知能力，即具身智能体能够很好地理解当前环境，如根据指令完成视觉定位、导航等任务。如今，基于MLLM的环境理解展现出极强的泛化能力，无须额外训练便可较好地理解环境，已在导航、视觉定位等应用中取得进展[20-22]。二是长程任务规划，即具身智能体能够将复杂指令分解为多个可执行的子任务。例如，指令“接杯水给我”被分解为多个步骤，包括找到杯子、拿起杯子、找到饮水机、将杯子放到接水口、打开饮水机、关闭饮水机、拿起杯子、找到用户、放下杯子等。“接杯水给我”这个指令可分为多个子任务，并且这个子任务的分解依赖于第一步的环境感知。在传统的具身智能体领域中，高级复杂任务的执行是一个难点。然而，大模型凭借其强大的逻辑推理能力，在解决此类问题上展现出显著的优势。目前，已有许多基于大模型的长程任务规划研究成果[23-24]。三是短程动作控制，即控制智能体的物理载体完成具体的子任务，如抓取、导航等。MLLM在实现短程动作控制上仍然面临着诸多挑战，通常通过调用API或生成可执行的代码来完成任务[25]。受益于MLLM的多模态数据处理能力，近年来出现了能够直接生成可执行动作的具身大模型[26-27]，这类模型可以接收并整合环境信息以及各类传感器信息等，并直接生成下一步可执行的动作信息。这些进展表明，MLLM有望进一步推动具身智能体在复杂现实场景中的应用。

随着MLLM的迅猛发展，具身智能体的研究也得到了极大推动。**图1**是以“embodied agent”为关键词的谷歌学术搜索结果，截至10月，2024年的发文量为1 350篇，斜线部分为按前10个月的平均发文量估计的后两个月的发文量。自2023年ChatGPT问世以来，具身智能体相关研究的数量显著增长。尽管已有部分研究对具身智能体领域进行了综述性回顾[28]，但这些工作并未涵盖基于多模态大模型的具身智能体研究。有部分回顾性工作虽然从整体上对具身智能领域进行了较为广泛的回顾[2,29]，但未深入探讨具身智能体的特定发展和挑战。为了弥补现有综述中的空白，帮助研究人员更迅速地掌握该领域的前沿进展，本文对基于多模态大模

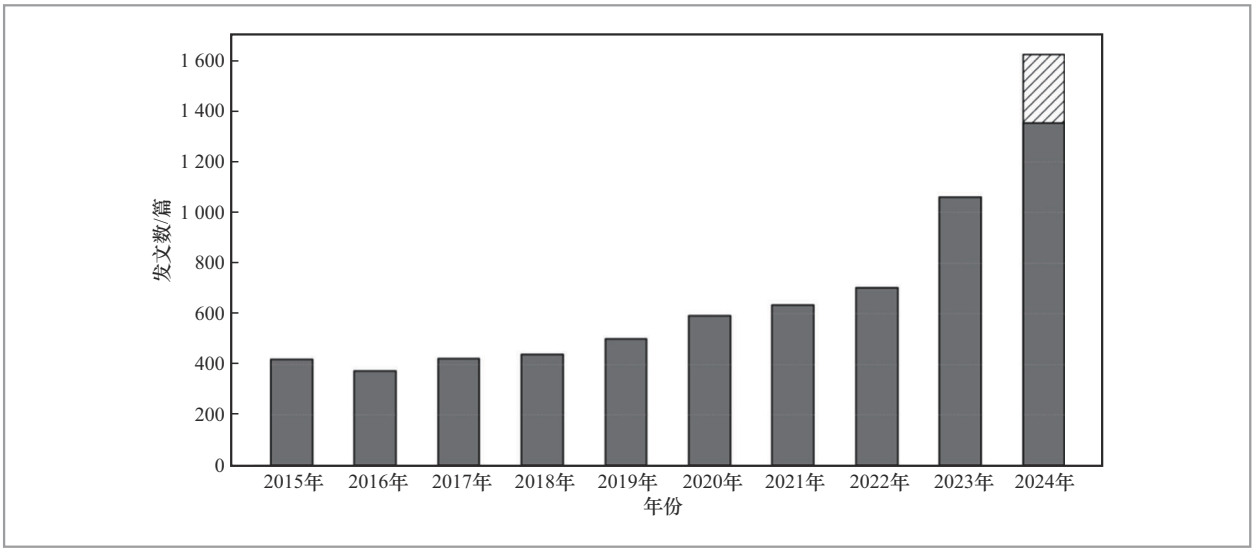


图 1 以“embodied agent”为关键词的谷歌学术搜索结果

型的具身智能体进行全面、系统的综述，基于多模态大模型的具身智能体综述的整体框架如图 2 所示。本文首先回顾并讨论了多模态大模型的发展现状；第二部分介绍了具身智能体领域的主要数据集；第三部分介绍了具身智能体的主要物理载体；第四部分展示了具身大模型领域的最新研究进展；第五部分探讨了具身智能体在高级任务规划中的应用；第六部分介绍了低级动作控制的相关研究；最后，总结了具身智能体目前面临的主要挑战，并对该领域未来的发展趋势进行了展望。

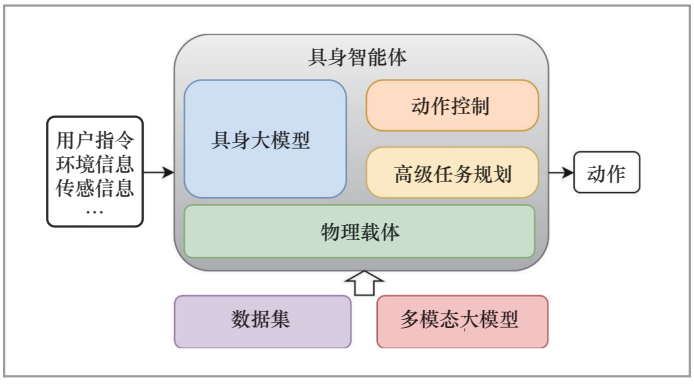


图 2 基于多模态大模型的具身智能体综述的整体框架

1 多模态大模型

在具身智能体中，视觉-语言模型是一种核心的多模态模型。本节重点介绍近年来视觉语言模型的发展脉络及具有代表性的研究进展。本节从视觉编码器、对齐方式、损失函数等方面对具有代表性的视觉语言模型进行了总结[30-38]，见表 1。

视觉语言模型是一类能够同时理解和处理视觉与语言信息的多模态模型。通过联合学习图像和文本特征，VLM 建立了视觉与语言之间的关联，被广泛应用于图像描述生成、视觉问答和图文检索等任务。Dosovitskiy 等[39]在 2020 年提出的视觉变换器（vision transformer，ViT）是视觉语言模型中一种常见的架构，成功超越了卷积神经网络等传统计算机视觉技术。该方法将二维图像划分为小块，将铺平后的一维向量映射到低维空间，再输入 Transformer 编码器中。2021 年，

表1 具有代表性的视觉语言模型

年份	模型	视觉编码器	对齐方式	损失函数	语言模型	训练数据
2021年	CLIP[16]	ViT/ResNet	对比学习	对比损失	/	图文对齐对比数据
2022年	Flamingo[30]	NFNet	感知重采样和门控交叉注意力	对比损失、LM loss	Chinchilla	图文交织数据集
2023年	BLIP2[31]	ViT	Q-Former+线性层	ITC loss、ITM loss、LM loss	FlanT5/OPT	可视化问答数据
2023年	Kosmos-1[32]	CLIP ViT-L/14	直接嵌入	LM loss	/	7 100万对图文交织文档数据
2023年	Kosmos-2[33]	CLIP ViT-L/14	直接嵌入	LM loss	/	GRIT数据集(9 100万张图像、1.15亿条对应文本、1.37亿个对应的包围框)
2023年	mini-GPT4[34]	ViT-G/14	Q-Former	LM loss	Vicuna	图文对齐数据,3 500对图文交织数据,3 500个额外高质量对话数据用于微调
2023年	QWen-VL[18]	ViT-bigG	位置敏感的视觉语言自适应器	LM loss	Qwen-7B	14亿图文数据对,35万条指令微调数据
2024年	LLaVA[35]	CLIP	线性层	LM loss	Vicuna	15.8万对对话指引集,5.8万条图文数据 COCO数据集
2024年	LLaVA-1.5[19]	CLIP-ViT-L-335px	多层感知机投影层	LM loss	Vicuna	55.8万对图文数据用于预训练,66.5万对数据用于微调
2024年	Instruct BLIP[36]	ViT	Q-Former+线性层	ITC loss、ITM loss、LM loss	FlanT5/OPT	26种对话任务的微调数据集
2024年	QWen2-VL[37]	ViT-bigG	位置敏感的视觉语言自适应器	LM loss	Qwens	1.4万亿tokens量的图文数据
2024年	MiniCPM-V[38]	SigLIP SoViT-400m/14	Llama3	LM loss	Llama3	约6.02亿条图文数据

OpenAI提出了对比语言-图像预训练(contrastive language - image pre-training，CLIP)模型，打破了固定标签的学习范式，实现了零样本分类任务[16]。CLIP的结构简洁优雅，通过编码器提取文本与图像特征，以配对样本为正样本、不配对的样本为负样本进行对比训练，从而在隐空间中对齐文本与图像的特征。2023年，Li等[31]提出了BLIP-2，其以较低计算成本实现了从视觉模型到大语言模型(large language model，LLM)的理解转换。具体而言，BLIP-2引入了查询变换器(querying transformer，Q-Former)，将视觉模型的输出作为自注意

力机制中的key和value，并结合可学习查询向量，利用交叉注意力融合文本信息。模型的训练分为两个阶段，先通过Q-Former将图像特征与文本特征对齐，再在冻结的LLM上进行训练，以生成与LLM对齐的特征。

微软与哥伦比亚大学合作开发的LLaVA通过GPT-4生成多模态的语言-图像指令跟随数据，并进行了指令调优，端到端地训练出了大型语言和视觉助手[35]。该方法首先利用GPT-4生成图像的指令对（比如“指令：总结一下这张图里面有哪些元素；回答：有<elements>”），并结合CLIP的图像特征，在线性投影之后与文本问题对应的特征拼接，交由LLM预测答案。LLaVA1.0以CLIP为图像编码器，以Vicuna为LLM，在后续推出的LLaVA1.5中使用CLIP-ViT-L-335px和多层感知机投影进行改进[19]。这些方法大多依赖图文配对的数据进行训练，这样的数据不易获得，并且这种方法可能会影响视觉元模型处理上下文的能力。2022年，Deepmind团队提出的Flamingo通过感知重采样技术和门控交叉注意力技术，将视觉多模态信息与LLM相结合，并使用图文交织数据进行训练[30]，即将图片直接嵌入上下文中进行训练。例如，“我有一只可爱的小狗，<image>，它正坐在凳子上”，该方式提升约17%的模型性能。KOSMOS-1基于Transformer架构采用语言模型作为通用接口，将多模态输入（如图像等）通过各自模态的编码器提取特征，与文本信息共同进行训练。该模型在大规模多模态语料库上进行端到端训练，语料库包括文本数据、图文匹配数据以及图文交织数据。此外，KOSMOS-1还通过纯语言数据进行指令遵循能力的微调。实验结果表明，KOSMOS-1在零样本和少样本学习设置下，能够很好地支持语言任务和感知任务，且表现出优越的语言-视觉理解能力[32]。KOSMOS-2在上个版本的基础上引入了图文基准数据，进一步提升了模型在图文基准和图文指代任务中的表现[33]。图文指代任务指的是模型能够理解指代词（如“图片中的黑衣男子”）所对应的图像实体，而图文基准任务则要求模型能够为不同的实体进行包围框的定位及校准。

OpenAI发布的GPT-4以其强大的视觉-语言能力被广泛关注[40]。由于OpenAI模型闭源，研究人员尝试通过开源模型和数据构建同等性能的多模态大模型，并推出了MiniGPT4[34]。MiniGPT4的架构与BLIP-2非常接近，采用了视觉编码器、Q-Former和LLM，只对投影层进行训练。由于低质量的图像文本数据会削弱LLM的性能，MiniGPT4采用了普通数据预训练和高质量数据微调的策略，最终取得了超越BLIP-2的性能。与此同时，国产的Qwen-VL系列视觉语言模型也达到了很高水平[18,41]。Qwen-VL与其他视觉语言模型类似，采用了预训练的Qwen-7B作为语言解码器，并使用了OpenCLIP预训练的ViT作为视觉编码器。在视觉编码器对齐过程中，Qwen-VL使用了位置敏感的视觉语言自适应模块，该模块仅包含一层交叉注意力，并使用可学习的查询向量进行特征交互与融合，以实现视觉信息与语言信息的有效对齐。Qwen-VL的训练在初期阶段固定LLM，在大规模数据上进行预训练，然后解冻LLM参数进行多任务训练，最后通过有监督的指令微调进一步提升模型性能。Qwen2-VL还引入了动态分辨率机制，通过动态调整不同分辨率下图像所使用的Token数量，结合多模态旋转位置嵌入，增强了模型的多模态处理推理能力，使其能够理解和建模更加复杂

的数据[37]。此外，面壁智能的MiniCPM系列在严格控制模型参数量的前提下，实现了可在边缘设备上部署的视觉语言模型[38]。GPT-4o、Gemini-Pro-1.5、Claude 3.5-Sonnet等也是非常优秀的闭源大模型，它们达到了业界先进水平[42]。

2 具身智能体领域主要数据集

具身智能体的训练与评测需要多模态、多任务的数据，涵盖视觉、语言、动作、环境交互等多种信息，这些数据是构建具身智能体的关键前提。近年来，很多高质量、大规模的数据集[43-50]涌现出来，见**表2**。

在机器人导航领域，Matterport3D[43]数据集提供了包含90个建筑场景的194 400张RGB-D图像和10 800个全景视图，并提供了表面重建、相机姿态及2D/3D语义分割标注。其精确的全局对齐和全景视图支持关键点匹配、视图重叠预测、语义分割等多种任务，为机器人导航与环境理解研究提供了重要支持。2019年，加利福尼亚大学伯克利分校等4个机构组织共同推出的一个大规模、多平台的机器人操作数据集RoboNet[44]，其涵盖了Sawyer、Kuka等多种机械臂在不同环境下的抓取、推动等任务数据。该数据集提供了丰富的视频、动作指令及传感器数据，为机器人操作、动作预测及模仿学习的研究提供了重要支持。为了实现更加通用的机械臂抓取，Dex-Net数据集[45]专注于抓取任务，提供了大量物体模型及其对应的抓取策略的相关数据。通过3D物体模型、抓取点标注及抓取成功率数据，该数据集为机器人抓取与抓取规划的研究提供了坚

表2 具身智能体相关数据集

数据集	年份	场景数	数据量	数据类型	适用任务	数据获取
Matterport3D[43]	2018年	90	1.08万	图像、3D网格、相机轨迹、语义标签等	室内导航、场景理解	真实
RoboNet[44]	2019年	113	1.62万	运动轨迹对应视频等	机械臂抓取、推动等	真实
Dex-Net[45]	2019年	/	500万+	点云、运动轨迹、力反馈传感数据、视觉数据等	机械臂抓取	仿真
Ego4D[46]	2021年	30+	3 000+	第一人称视频、深度信息等	行为理解、物体操控等	真实
ALFRED[47]	2019年	/	2.5万+	自然语言指令、动作轨迹、环境信息等	机器人在家居环境中执行基于自然语言指令的各种任务	仿真
Open X-Embodiment[48]	2023年	/	200万+	运动信息、多模态感知信息等	通用机械臂操作	真实、仿真、开源数据集
RoboMIND[49]	2024年	/	5.5万	视觉信息、机器人感知数据、指令等	机械臂、灵巧手等	真实
ARIO[50]	2024年	/	300万+	运动信息、多模态感知信息等	多类型机器人训练	真实、仿真、开源数据集
AgiBot World	2024年	/	100万+	运动信息、多模态感知信息等	通用机械臂操作	真实

实的基础。在视觉理解方面，Meta AI联合多个学术机构共同构建了Ego4D[46]，这是一个大规模的第一人称视角的视频数据集，旨在推动以自我为中心的视频理解和具身智能的研究。Ego4D涵盖了丰富的日常活动、社交互动和任务执行场景，为具身智能体的视觉理解能力提升提供了重要数据支持。此外，为研究机器人任务执行而设计的ALFRED数据集[47]，主要用于训练机器人在复杂环境中执行基于自然语言指令的各种任务。数据集的任务主要是在虚拟环境中执行日常任务，任务指令通常以自然语言的形式呈现。

自2023年以来，随着具身智能领域的快速发展，一系列更大规模、更高质量的数据集涌现出来，为通用机器人策略模型的训练与评测提供了重要支持。其中，Open X-Embodiment[48]是一个跨机器人平台、跨任务、跨环境的大规模数据集，支持训练通用的机器人策略模型（如RT-X模型）。该数据集涵盖了22台机器人、527种技能、160 266个任务以及200万条任务片段，包含多种环境和场景的数据，为机器人领域的通用化研究奠定了坚实基础。与此同时，国内研究团队在这一领域也取得了显著进展。例如，RoboMIND[49]数据集由国家地方共建具身智能机器人创新中心等团队提出，包含5万多条任务轨迹和279个不同任务，支持机械臂、灵巧手等多种机器人的训练，为具身智能体的开发提供了多样化任务场景。此外，鹏城实验室等机构提出的ARIO数据集[50]从258个系列中收集了约300万条任务片段和321 064个任务，旨在为机器人抓取、操作和任务规划等研究提供高质量的数据支持。为了进一步支持更泛化、通用的具身大模型训练，智元机器人等国内机构推出了AgiBot World数据集，其包含上百万条任务轨迹，长程数据的规模更大，为复杂任务和多样化环境的适应提供了重要资源。这些数据集的涌现不仅推动了具身智能领域的技术进步，也为通用人工智能机器人的发展提供了强有力的数据支撑。

3 物理载体

与传统人工智能系统不同，具身智能体需要通过其具象化的物理形态与环境进行交互，从而完成预定的任务。不同物理载体的选择不仅影响具身智能体的形态，还决定了其适用领域与功能扩展。常见的具身智能体物理载体包括机械臂、灵巧手、轮式无人车、无人机、四足机器人、人形机器人等，如**图3**所示。**图3（a）**为Universal Robots公司推出的UR20机械臂，**图3（b）**为傲意科技推出的ROH-A001灵巧手，**图3（c）**为申昊科技推出的SHIR3002轮式无人车，**图3（d）**为大疆推出的DJI Mavic3 Pro无人机，**图3（e）**为宇树科技推出的A1四足机器人，**图3（f）**为智元推出的远征A2人形机器人。

3.1 机械臂

机械臂是目前使用最广泛的机器人之一，被广泛用于工业制造、医疗服务等领域[51]。机械臂由多个关节和执行器组成，能够执行一系列复杂操作，如抓取、焊接、喷涂等。其核心功能是通过关节的运动来控制末端执行器的位置与姿态，这一过程依赖于正运动学和逆运动学的计算来确定各关节的运动量[52]。基于动力学理论建立的数学模型（如牛顿-欧拉方程和拉格朗日方程），能够准确描述机械臂的运动状态及其受力关系[53]。通过精密的控制系统，

图 3 常见的具身智能体物理载体

机械臂可以实现位置、速度和力的控制，从而顺利完成特定任务。尽管机械臂在许多应用场景中表现出色，但其通常具有较低的自由度，影响了在高度复杂任务中的效果。此外，大多数机械臂为固定设备，限制了其操作的灵活性和活动范围。然而，由于机械臂通常具备较高的操作精度、相对低廉的成本以及较易编程的特点，机械臂成为具身智能体的理想物理载体之一。近年来，研究者已将多模态大模型引入机械臂的抓取任务中，这不仅增强了机械臂与用户交互的能力，还显著提高了其在多样化任务中的泛化能力[24, 54]。

3.2 灵巧手

灵巧手是一种仿生机械手，旨在模拟人类手部精细运动，从而实现复杂的抓取和操控任务。灵巧手通常由5根手指组成，每根手指配备多个独立的驱动器，拥有多个自由度。在常见设计中，整个灵巧手通常拥有多达24个自由度，以更好地模仿人手的灵活性和功能。灵巧手的关节和手指具备高度灵活性，并且通过触觉、力觉和视觉传感器感知物体特性及环境变化。这些传感能力为灵巧手在工业自动化、医疗辅助和服务机器人等领域的应用提供了基础，尤其适用于执行精密组装、微创手术及日常家务等任务。相较于机械臂，灵巧手的复杂性远超机械臂，其精确控制和智能化操作也更加具有挑战性。由于灵巧手具有高自由度，传统的基于规划的动作控制方法往往难以应对复杂操作[55]，近年来，随着强化学习技术的发展，强化学习逐渐成为灵巧手控制的主流方法[56-57]。研究者通过强化学习训练灵巧手，以应对传统方法中对真实数据依赖较大的问题。此外，随着MLLM的兴起，研究者逐步将MLLM应用于灵巧手的任务导向控制。部分研究成功利用MLLM实现了灵巧手对未

知物体的抓取任务[58]，该方法依赖MLLM对环境中物体位置的感知能力。此外，有研究通过MLLM生成了基于自然语言指令的灵巧手抓取数据集，并提出了一套指令对齐的抓取动作生成框架[59]，实现了灵巧手根据指令执行多样化抓取任务的能力。还有研究将人类经验整合进MLLM，以提升灵巧手的抓取成功率[60]。基于MLLM的灵巧手具身智能体展现出广阔的前景，但当前仍面临诸多挑战，如缺乏真实数据集、如何设计更高效的动作生成网络以及如何更好地融合MLLM等问题。

3.3 轮式无人车

轮式无人车是一种基于轮式移动平台实现移动的自主或遥控无人车，因其结构简单、能效高且易于控制，被广泛应用于工业、服务业、农业以及科研教育等领域。由于轮式移动方式产生的摩擦较小，轮式机器人通过调整车轮的转速和方向来实现在平坦地面上的高效、稳定移动。轮式机器人常见的设计包括两轮差速、四轮驱动或全向轮（麦克纳姆轮）系统，以实现前进、后退及复杂转向等多样化动作[61]。相比于其他移动方式（如履带式），轮式无人车具备低能耗和高导航精度的优势，特别适合用于工业自动化中的物料搬运、服务机器人中的自主导航以及安防巡逻等任务[62]。然而，轮式无人车在复杂或恶劣环境下的机动能力较为受限，其承载能力也有限。此外，目前大多数轮式无人车主要被应用于在已知环境中执行特定任务。然而，随着多模态大模型的迅猛发展，轮式无人车作为具身智能体的潜力逐渐显现。借助MLLM，轮式无人车具备更强的环境感知、自主决策和任务理解能力，从而能够更好地理解用户意图、探索未知环境并执行复杂任务[63-64]。

3.4 无人机

无人机是一种无须人工驾驶的飞行器，被广泛应用于农业、物流、摄影、军事和环境监测等领域[65]。无人机通常由机体、动力系统、控制系统、传感器和通信系统组成，利用电池供电的马达提供推力，并通过飞行控制器和导航系统维持飞行的稳定性[66]。卓越的灵活性使无人机能够到达人难以进入的区域，并快速采集实时数据，大幅提高任务决策的效率。在无人机的导航、避障及目标识别等方面，人工智能算法的应用已经相当普遍[67]。然而，传统的AI算法在无人机的应用上存在一些限制，特别是在高度动态或极端环境下。具体来讲，这些算法对结构化数据的依赖以及预设算法的局限性影响了无人机的表现。此外，现有的无人机系统无法执行较为复杂的任务，无法进行自然语言的交互，缺乏高水平的自主决策能力。而以无人机为载体的具身智能体有望突破这些限制，通过自然语言交互及对环境信息的感知实现更高级别的自主决策，并自主探索未知环境[68]。这一发展为无人机在更多复杂场景中的应用提供了无限可能。

3.5 四足机器人

四足机器人模拟四足动物的运动模式，具备4条独立驱动的机械腿，每条腿由多个关节和驱动器组成，能够实现复杂的步态规划与动态平衡控制。相较于轮式机器人，四足机器人在复杂和崎岖地形中的通过性更强，适用于各种地面条件（如草地、砂石、山地和楼梯等），展现出极高的环境适应性。四足机器人的运动方式多样，包括小跑、奔跑、跳跃等，并通过精密控制

算法实时调整步态，以保持平衡和灵活运动。其感知系统通常配备激光雷达和摄像头，以增强环境感知能力，从而实现自主导航及障碍物规避[69]。四足机器人由于其卓越的稳定性和灵活性，被广泛应用于工业巡检和救援任务。四足机器人的高负载能力和出色的环境适应性使其能够在恶劣环境中执行复杂任务[70]。例如，四足机器人已被应用于工业检测和维护工作，并能够配备自主导航和远程操作功能，胜任长时间的户外任务，甚至适用于极端环境中的任务（如月球任务[71]）。尽管四足机器人在性能和适应性方面表现优异，但其复杂设计和高昂的制造成本使得初期投资较大。此外，其在复杂环境中的电池续航能力较为有限，通常需要频繁充电或更换电池[72]。四足机器人具备优异的稳定性和对复杂环境的适应性，成为具身智能体的极具潜力的物理载体[73]。

3.6 人形机器人

人形机器人是一类仿照人类外形和运动方式设计的机器人，通常具备与人类类似的躯干、手臂、腿和头部结构，目的是通过模仿人类的动作实现与人类社会的自然交互。作为机器人领域的集大成者，人形机器人往往结合了包括灵巧手在内的多种先进技术。由于拥有与人类相似的肢体结构，人形机器人能够行走、奔跑、上下楼梯、搬运物品等，适用于在专为人类设计的环境中执行各种任务。人形机器人通常采用多关节驱动与精确的控制算法以完成复杂的运动任务。例如，波士顿动力公司发布的Atlas机器人展示了极强的运动协调能力，能够完成跑步、蹦跳等复杂动作[74]。特斯拉和宇树科技推出的机器人使用电机驱动系统取代传统的液压驱动，在保证良好的运动能力的前提下降低了制造成本，为人形机器人进入市场奠定了基础。软银公司的Pepper机器人侧重于情感识别与自然语言交互，提升了用户的互动体验[75]。FIGURE公司推出的Figure 02机器人结合了大模型技术，具备更强的视觉感知和逻辑推理能力。以人形机器人为载体的具身智能体被认为是具身智能体发展的终极目标之一。随着MLLM的发展，这一目标正在逐步实现[76]。

4 具身大模型研究进展

具身智能体的核心能力之一是具备视觉感知、指令理解以及自我感知的能力，能够有效执行复杂的交互和动作策略。近年来，VLM的发展为具身智能体的视觉感知能力提供了有力支持，同时其语言理解和推理能力也使智能体能够准确理解指令。然而，VLM在感知自身状态以及生成可执行的动作指令方面存在明显局限性，无法直接支持复杂的交互和操作任务。为了解决这一问题，具身大模型应运而生，逐渐成为具身智能体的基座。具身大模型不仅继承了多模态大模型的环境感知与指令理解能力，还能够感知自身状态并生成具体的动作策略，从而实现可执行的动作策略。具身大模型近年来取得了显著进展，催生了一系列优秀的研究工作[77-86]，见**表3**。

2022年，谷歌团队提出了RT-1模型，将图像和指令信息映射到同一空间，并使用Transformer架构进行端到端训练，直接估计智能体的动作策略[26]。RT-1模型通过FiLM[87]编码图像信息，并利用TokenLearner[88]编码指令信息，成功实现了对操作机械臂、基座、夹爪等系统的

表 3 具有代表性的具身大模型

年份	模型	视觉编码器	策略头	物理载体	模型大小	训练数据	控制频率	输出
2022年	RT-1[26]	FiLM-EfficientNet	直接预测	可移动机械臂平台	35 MB	1.3亿条	3 Hz;10 Hz	动作参数
2022年	RoboFlamingo[77]	ViT	LSTM	六轴机械臂	3 B/4 B/9 B	1000条(CALVIN)	—	动作参数
2023年	RT-2[78]	ViT	直接预测	可移动机械臂平台	5 B/12 B/55 B	13万 条	1~3 Hz (55 B);5 Hz(5 B)	动作偏移对应的token
2023年	RT-X[48]	同RT-2/1	同RT-2/1	可移动机械臂平台	同RT-2/1	200万+条	同RT-2/1	同RT-2/1
2023年	GR-1[79]	ViT	多层感知机	六轴机械臂	195 MB	1000条(CALVIN)	—	下一步视频及对应动作参数
2023年	SayCan[23]	—	基于打分筛选	可移动机械臂平台	540 B	27.6万条	—	动作参数
2023年	Q-Transformer [80]	FiLM-EfficientNet	基于Q-value选择	可移动机械臂平台	35 MB	3.8万条	3 Hz	Q-value
2023年	PaLM-E[27]	ViT	—	—	12 B/84 B/562 B	—	—	子指令
2024年	RT-H[81]	ViT	直接预测	可移动机械臂平台	55 B	10万条	—	子任务文本及对应动作偏移token
2024年	EmbodiedGPT[82]	ViT-G/14	多层感知机	可移动机械臂平台	10 B	2 927小时 视频(EgoCOT)	—	子指令及对应的动作参数
2024年	GR-2[83]	VQGAN	多层感知机	六轴机械臂	30 MB/95 MB/312 MB/719 MB	5 000条	—	下一步视频及对应动作参数
2023年	RT-Trajectory[84]	FiLM-EfficientNet	直接预测	可移动机械臂平台	35 MB	7.3万条(2D轨迹数据)	3 Hz	动作参数
2024年	SARA-RT[85]	sViT	直接预测	机械臂	5 B	仿真器	10 Hz	动作向量表征
2024年	RoboMamba [86]	CLIP/SigLIP ViT-L	多层感知机	机械臂	2.7 B	仿真器	—	动作矩阵

逐步动作位置估计。RT-2模型在RT-1的基础上，通过引入预训练的PaLM-E[27]模型显著增强了模型的泛化能力，同时保持了RT-1的端到端动作估计能力[78]。RT-X在RT-2的基础上进一步扩展，在大规模的Open X-Embodiment数据集上进行训练，进一步提升了模型的性能和效果[48]。此外，RT-H模型通过层级查询的方式串行执行任务，首先生成子指令，然后基于子指令生成可执行的动作策略。

RT-H 的独特之处在于子指令和动作策略的生成在同一个具身大模型上完成，这使得任务执行更加高效和灵活[81]。

具身大模型的结构与人脑在执行任务时的信号处理方式具有高度相似性：首先，通过高级皮层区域处理高级抽象任务；然后，信号被传递至执行具体任务的低级区域（如丘脑、脑干、小脑）；最后，通过脊髓将信号传递至运动器官执行动作[89]。近年来，研究人员逐渐减少了对大模型直接估计动作策略的依赖，而是通过在大模型后添加策略头来输出动作信息，这一设计已成为主流的具身大模型框架（如图 4 所示）。该具身大模型框架需要根据用户的指令、当前的机器状态以及环境信息预测下一时间点的动作参数。其通常可通过多模态的编码器将视觉、文本、机器状态信息映射到同一空间，并通过大语言模型输出特征，再将输出的特征经过策略头输出机器的动作参数。策略头通常是一个简单的网络，有研究采用全连接网络，也有研究采用长短程记忆网络[77, 79, 82]。EmbodiedGPT 提出了 EgoCOT 和 EgoVQA 数据集，该数据集基于用户第一视角的视频和简单注释，利用 ChatGPT 生成精细的子指令。基于此数据集，笔者训练了一个多模态大模型，能够基于视频实现精确的动作规划，并生成相应的动作指令[82]。此外，RoboFlamingo 验证了具身大模型无须从头训练，直接微调预训练的 VLM 也能获得良好的表现。在该研究中，作者以长短程记忆网络为策略头，进一步提升了效果[77]。GR-1 使用了大规模视频-文本数据进行预训练，自回归生成

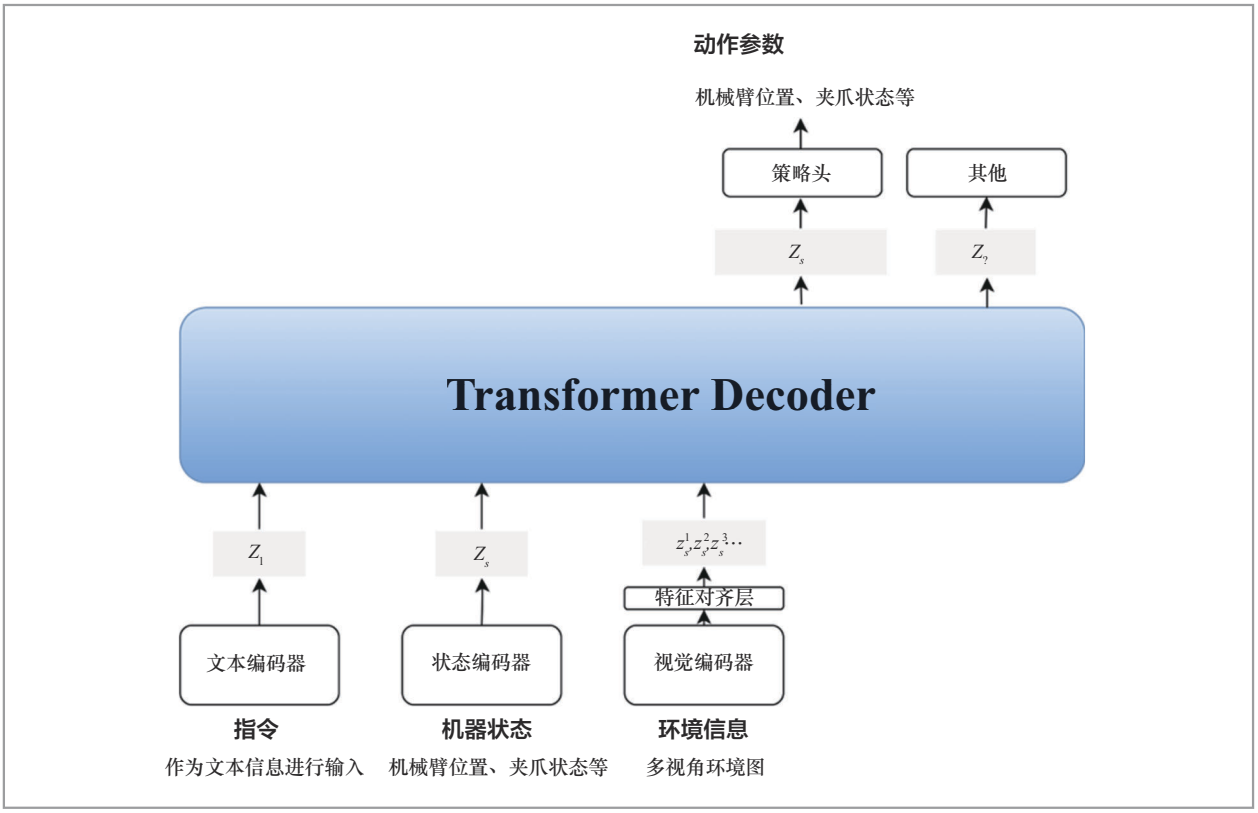


图 4 常见的具身大模型框架

下一步视频，然后通过微调训练出动作估计策略，结果表明视频-文本数据预训练显著提升了模型的性能[79]。GR-1的策略头采用了多层感知机，GR-2在GR-1的基础上扩展了训练数据规模，取得了更佳的效果[83]。

除了上述框架，还有研究探索了如何将强化学习与具身大模型相结合。例如，SayCan模型使用LLM进行长程任务分解，然后使用强化学习中的Q-learning算法训练价值函数，以判断子任务的可行性，最后利用行为复制生成相应动作[23]。Q-Transformer则提出了一种将Transformer架构与Q-learning相结合的方式，设计了保守正则项，通过优化TD误差和正则项训练模型，并使用蒙特卡洛法加速学习过程[80]。

具身大模型的训练通常依赖大量数据，然而收集机器人的动作数据成本较高，为此，研究者提出了多种应对数据不足的方法。例如，AutoRT团队借助有限的人为监督，运用VLM进行导航，并利用LLM生成安全的任务和可执行的运动策略。通过这一策略，他们在4栋建筑中通过20台机器人收集了大量多样的真实数据[90]。为了加快具身大模型的推理速度，SARA-RT提出了线性复杂度的自注意力机制，极大提升了推理效率[85]。此外，RT-Trajectory[84]通过引入RGB轨迹图作为提示，提升了动作的泛化能力，显著提高了任务成功率。

总体而言，具身大模型通过端到端的动作策略输出方式，为具身智能体的实现提供了一个可行的技术路径。然而，训练此类大模型往往需要大量的数据，而收集机器人操作数据需要高昂的成本，当前大多数具身大模型主要聚焦于机械臂等少自由度的操作领域，且现有的模型在推理速度方面尚未达到流畅的动作执行效果。

5 高级任务规划

人类面对复杂任务时，通常会对任务进行分解，逐步完成子任务以实现最终目标。同样的解决方案也适用于具身智能体。当指令变得非常复杂时，仅依赖一个简单的策略网络难以完成整个任务。通常的做法是将复杂的高级任务分解为多个子任务，并确保这些子任务能够被顺利执行。例如，“帮我倒杯水”这样的高级任务可以被分解为“找到杯子”“找到水壶”“拿起水壶”“移动至杯子”“倒水”等多个子任务。此过程涉及多种能力，如“找到杯子”需要具身智能体具备充分的环境感知能力，若在当前视野中未找到目标，还需主动探索。而“拿起水壶”等任务涉及机械臂抓取操作。具身大模型可以直接生成这些子任务执行的动作策略[26,79]，此外还可以预先定义好子任务执行的工具或其他辅助元素[24-25]。本节将介绍如何实现高级任务的分解。

传统的高级任务规划依赖符号规划方法（如STRIPS[91]、PDDL[92]）和搜索算法（如A*[93]、MCTS[94]），通过预定义的规则和启发式算法生成任务步骤。这些方法在结构化环境中表现较好，但在应对动态变化和复杂场景时存在一定的局限性。为提升机器人在复杂环境中的自主性和适应能力，研究者引入了多模态大模型（如视觉语言模型）来执行具身任务规划。与传统方法不同，多模态大模型结合了视觉、语言和其他感知数据，能够实现从感知、理解到执行任务的全流程能力。通过引入链式思维等推理方法[95]，多模态大模型能够将高层次的任务指令分解为一系列逻辑

步骤，生成合理的行动计划。这使得模型能够适应动态变化的环境，而不再依赖传统的预定义规则。通过引入多模态大模型，具身任务规划逐步从依赖固定规则的逻辑推理转向了基于感知与语言理解的智能推理模式。这一转变使得机器人在处理复杂、多变的环境时，表现出更强的智能性和灵活性。

多模态大模型在高级任务规划中展现出诸多优势，但仍面临一些挑战。例如：子任务的分解必须在现实世界中可行，这要求大模型生成的子任务能够在可执行的范围内进行；大模型在处理长任务序列时可能会遇到记忆问题或信息丢失，导致错误的任务推断或步骤遗漏；大模型的幻觉问题也可能导致分解出的子任务难以执行，或者无法实现最初的指令目标[96]。为了更好地利用大模型来进行高级任务规划，研究人员从不同方面进行了探索[97-112]，如**图5**及**表4**所示。

5.1 提示工程

大模型具有通用的语义理解能力，但对输入格式和提示的敏感度极高。通过精心设计提示词，用户可以提升模型在特定任务中的表现，并减少生成无关或无意义内容的可能性。提示工程正是利用这一点，通过反复试验和优化设计出最佳的输入形式，以获得更优质的输出。鉴于大模型在代码生成方面的卓越表现，已有研究将复杂任务的分解转化为代码生成问题[24, 25, 97]。例如，CaP首先定义了一些具有特定功能的可执行函数，然后将这些函数嵌入提示中，通过提供示例来引导大模型生成代码以完成指令[24]。PROGPROMPT[97]同样通过提示生成可执行代码，但在提示中增加了环境中可用的操作和对象的类似规范，从而提高了生成操作的可行性。Instruct2Act[25]通过定义环境感知的函数（利用SAM对环境中的实体进行分割，并使用CLIP获取其语义信

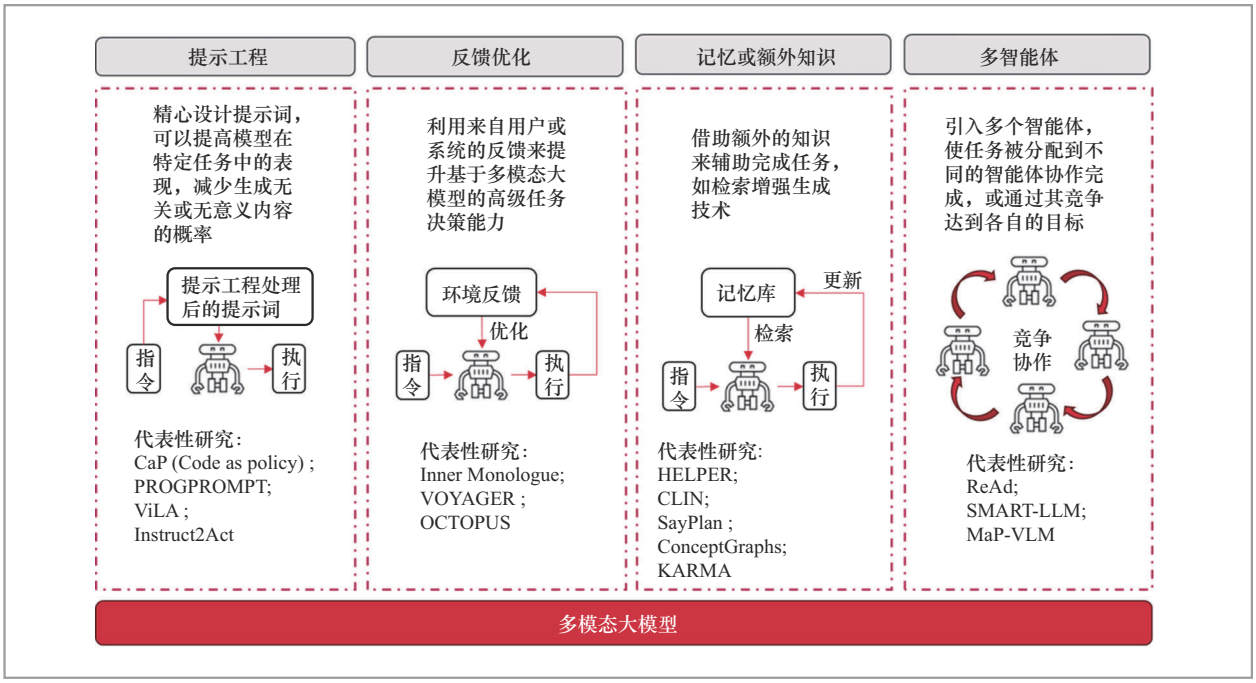


图5 主要的基于多模态大模型的高级任务规划方法

表 4 基于多模态大模型的高级任务规划

类型	年份	方法	基本思想	大模型	评估数据集
提示工程	2022年	CaP[24]	定义一些具备一定功能的可执行函数，然后将其编排在提示中，并通过一些示例来提示大模型直接生成代码来完成指令	GPT-3	HumanEval
	2022年	PROGPROMPT [97]	用环境中可用操作和对象的类似程序的规范，以及可以执行的示例程序来提示LLM	GPT-3	VirtualHome
	2022年	Translated-LM[98]	将大模型分解的子任务校正为可执行的格式	GPT-3	VirtualHome
	2022年	Reflexion [99]	通过反思的方式发现不正确的步骤并进行修改	GPT-4	ALFWorld
	2023年	Instruct2Act[25]	在CaP的基础上使用CLIP和SAM来实现环境感知	GPT-3.5	VIMABench
	2023年	ViLA[100]	通过思维链推理来分解子任务，同时在提示中使用了视觉输入	GPT-4V	RAVENS
	2024年	RSFT[101]	针对需要经过一定思考的高级任务，设置提示来分解子任务	GPT-3.5	VIMABench
反馈优化	2022年	Inner Mono - logue[102]	将不同类型的反馈(包括环境、用户交互、执行是否成功等)输入LLM，形成闭环的任务规划	InstructGPT	RAVENS
	2023年	VOYAGER[103]	通过环境反馈、执行结果持续扩展技能库，实现LLM驱动的终身学习的具身智能体	GPT-4	《我的世界》
	2023年	OCTOPUS[104]	利用环境反馈的强化学习微调视觉语言模型，使其能够实现更加精准的任务规划	MPT-7B	OctoGibson
记忆或额外知识	2023年	HELPER[105]	将相似任务经历、对话、场景描述等存储到记忆库中，利用检索增强技术提升LLM对高级任务的分解执行能力	GPT-4	TEACh
	2023年	CLIN[106]	通过因果总结技术，基于LLM反思任务执行失败的原因，并将反思出的知识存储在记忆中，并辅助后续的类似任务的执行	GPT-4	ScienceWorld
	2023年	SayPlan [107]	构建3D场景图，借助LLM进行语义搜索，然后结合经典的路径规划算法，并根据使用场景图模拟器反馈的结果进行修正	GPT-3.5/4	Office/Home
	2023年	ConceptGraphs [108]	通过利用2D基础模型的输出，然后结合多视角关联，将其转换到3D表示中，进而构建出3D词汇图，从而实现复杂任务的高级语义推理和规划	GPT-4	Replica
	2024年	KARMA [109]	使用长程记忆来构建3D场景图，使用短程记忆来动态记录目标的位置状态信息。LLM通过查询记忆来实现更好的高级任务规划	GPT-4o	AI2-THOR
多智能体	2024年	ReAD[110]	使用多智能体强化学习的方式学习一个奖励函数来判断单个智能体及其要执行的子任务打分	LLama-3.1-70B-Instruct	OvercookedAI 和 RoCoBench
	2024年	SMART-LLM[111]	通过LLM赋予不同机器人不同的能力，然后将分解任务并分配给不同机器人协同执行	GPT-4	AI2-THOR
	2024年	MaP-VLM[112]	使用3个智能体分别收集环境信息、任务规划以及评估规划可行性	GPT-4V	ALFRED

息），结合大模型生成包含环境感知能力的可执行代码，最终实现了机械臂的操作。此外，Translated-LM[98]研究发现，大模型在高级任务规划时生成的子任务虽然在语义上合理，但往往无法直接执行。例如，若任务是“刮胡子”，大模型可能生成“我要首先抓取刮胡刀……”这样的子任务，但实际应为“抓取刮胡刀”。研究者将生成结果与可执行动作库进行对比，从而对生成结果进行校正。在RSFT中，研究者通过在提示词中设定输出格式并结合少量示例提示的方式来分解子任务[101]。

为了进一步提高生成结果的可靠性，一些大模型推理提示技术也被应用于具身智能体的高级任务分解。在ViLA中，研究者使用思维链推理方法，提升了每一步子指令的可行性，同时将视觉信息融入提示中，使模型在任务分解时能够充分考虑环境因素[100]。Reflexion[99]通过对生成任务的反思发现分解子任务中的错误，并在后续执行中对错误的规划进行修正。上述提示工程的方法显著提升了基于大模型的具身智能体在高级任务规划中的表现。

5.2 反馈优化

不同于其他类型的智能体，具身智能体强调与环境和用户的交互，在此过程中能够接收到来自环境或用户的反馈。如何利用这些反馈信息提升具身智能体的高级任务规划能力，从而实现闭环系统，是该领域的重要研究方向之一。Inner Monologue[102]是早期将环境反馈引入大模型驱动的具身任务规划的研究之一，其构建了一套闭环的任务规划系统，将环境观测、用户反馈以及执行结果都作为反馈输入大模型，进而促使大模型进行二次规划以完成复杂任务。有研究通过引导大模型向用户询问提示中缺失或模糊的信息，以用户反馈为依据，生成更准确的任务规划结果[113]。VOYAGER[103]则通过环境反馈和执行结果持续扩展自身的技能库，随着执行任务的增多，智能体的技能库愈加丰富，从而实现了基于LLM的终身学习具身智能体。与上述方法不同，OCTOPUS[104]直接利用环境反馈信息，通过强化学习方法对视觉语言模型进行微调，从而提升其任务规划能力。此外，其他研究（如ReAD[110]、PaLM-E[27]、CLIN等）也充分利用反馈信息，进一步优化模型在复杂环境下的表现。

5.3 记忆或额外知识

人类大脑在处理复杂任务时，通常会依赖以往的经验，并借助额外的知识辅助任务的完成。在基础大模型的研究中，研究人员也运用了这一基本原理。例如，检索增强生成技术已被广泛应用于各类大语言模型，以缓解模型本身产生的幻觉，从而生成更可靠的结果[114]。在具身智能体的高级任务规划中，利用智能体自身的记忆或额外的知识同样能够提升任务规划的准确性。HELPER[105]系统将相似任务的经历、对话记录和场景描述等信息存储在记忆库中，在任务规划过程中检索相关数据，并将其加入提示词中，从而提升LLM在高级任务分解和执行上的效果。此外，CLIN[106]通过因果总结技术，基于大模型对任务执行失败的反思，提炼出失败的原因并将其存储在记忆中，为后续类似任务提供辅助。

除了利用智能体的任务执行经验外，研究者还可以通过事先构建的额外知识来增强任务规划能力。SayPlan[107]提前构建环境中对象的3D场景图，智能体在任务规

划时借助LLM搜索相关目标，随后结合经典的路径规划算法，基于3D场景图进行路径规划分析。在此基础上，智能体通过在场景图中模拟执行的方式，对初步生成的任务规划进行修正。类似地，ConceptGraphs将智能体的任务执行场景构建为大模型可读的3D词汇图。它先获取2D基础模型的输出，再通过多视角关联将其转化为3D表示，最终生成用于处理高级复杂任务规划的3D词汇图[108]。为了加快对3D场景图的更新，KARMA [109]采用长程记忆来构建3D场景，并使用短程记忆动态记录目标的位置和状态，而大模型通过查询这些记忆实现更精确的高级任务规划。

无论是利用先前执行任务的对话记录、从失败中总结出的经验，还是通过事先构建的环境信息，记忆或额外知识的引入显著增强了具身智能体的任务规划能力。然而，如何构建有效的知识库、如何动态更新现有的知识库，以及如何从过往经历中提取出有价值的经验，仍然面临诸多挑战。

5.4 多智能体

多智能体技术引入多个智能体，使复杂任务能够被分解并分配给不同智能体协作完成，或者在同一环境中竞争以达到各自的目标。有研究者提出为3个智能体分别赋予不同的能力：感知智能体负责收集环境信息(基于视觉语言模型)；任务规划智能体负责制定计划（基于LLM)；评估智能体评估规划的可行性。3个智能体相互协作，大幅提升了任务成功率[112]。此外，SMART-LLM[111]在更低的层级为不同智能体分配不同的职责，如负责目标搜寻、负责具体操作，多智能体相互协作完成任务。

不同于上述方法直接使用预训练的大模型，ReAD方法首先通过大模型生成各种子任务，然后通过多智能体强化学习的方式训练一个奖励函数，用于对单个智能体及其执行的子任务进行评分。在具体执行之前，奖励函数会评估任务分解的效果及子任务的可行性，从而提高复杂任务执行的整体成功率[110]。

然而，使用多智能体来完成高级任务规划，虽然可以提升任务执行的成功率，但是增加了计算量。此外，如何实现智能体之间的协调控制、信息共享以及动态环境适应存在一定挑战。

6 动作控制

通过高级任务规划将复杂任务分解为可执行的子任务后，智能体需要控制其物理载体完成这些任务，这便涉及动作控制。动作控制的核心目标是制定合适的运动策略，以完成一系列低层次的动作。例如，“抓取杯子”便是一个典型的动作控制任务。具身智能体的物理载体类型各异，其动作控制的难度也不相同，因此产生了多种控制策略。基于强化学习的动作控制方法是较为常见的选择[115]，同时也有研究采用模仿学习直接生成可执行动作[116]。本节将重点介绍几类常见的物理载体的动作控制，包括机械臂动作、双足行走和灵巧手的控制策略。

6.1 机械臂动作控制

机械臂的动作控制旨在通过精确调控多个关节的运动，实现对物体的抓取、移动和操作。例如：具备6个自由度的机械臂只需获取末端执行器的目标位置和姿态(通常由3个空间位置参数和3个旋转参数构成)，便可通过逆向运动学计算出各个关

节的角度及运动路径。因此，机械臂的动作控制问题可以转化为对这些参数的精确估计。有研究通过预先设定的工具完成机械臂的抓取等低级任务[23-25]。例如：Instruct2Act方法使用SAM[117]对环境中的主要物体进行分割，利用OpenCLIP[118]将环境物体与指令中的信息对齐，从而获取目标物体的位置信息，实现动作控制。然而，使用这些外部预定义工具可能会引入额外的延迟，影响实时任务的响应速度和效率。此外，系统的整体性能很大程度上依赖于外部预定义工具的质量，若效果不佳，即使高级任务规划合理，也可能导致智能体无法顺利完成任务。

与上述方法不同，一些研究将语言指令与机器人低层动作通过奖励函数进行连接，并利用大语言模型将任务描述转化为奖励函数（用代码表示）。这些奖励函数用于指导轨迹优化和强化学习算法，帮助机器人学习最优策略。结果表明，在机械臂的复杂操作任务中，这种结合大语言模型与强化学习的方法能够显著提升任务完成效果，在复杂机器人控制中展现出巨大的应用潜力[119]。此外，还有一些研究者提出了基于强化学习的具身智能基础模型，并在多种应用场景中取得了优异的表现[120]。

除了使用强化学习策略外，还有研究采用模仿学习方法训练模型，使其能够直接输出低层次的动作策略，如RT-1[26]。

6.2 双足行走控制

双足行走控制涉及通过精确调控机器人的双足运动，实现平稳行走和姿态平衡的过程。相比于机械臂控制，这一任务的复杂性更高，难以通过端到端的方式直接训练出能够实现双足行走的具身大模型。传统的双足机器人大多依赖于基于动力学模型的力矩控制技术，如虚拟模型控制[121]。基于动力学模型的力矩控制方法通过简化机器人的物理模型，并施加虚拟力进行控制，在保证控制效果的同时，也能在一定程度上减少计算量。此外，双足行走控制还需精心设计步长和步频等运动参数，以确保运动的稳定性。

近年来，深度强化学习在双足行走控制中的应用逐渐成为主流[122]。研究人员通常在仿真环境（MuJoCo [123]和Isaac Gym [124]）中训练双足控制策略，并将成功的策略迁移至现实世界。有研究专注于逆强化学习，Wu等[125]提出了一种新算法，能够在复杂地形上实现双足行走，并通过学习专家奖励函数优化行为策略。也有研究聚焦于把机器人系统知识与强化学习相结合，着重开展以脚部设定点为基础的任务空间动作的学习[126]。此方法结合了任务空间策略与基于模型的逆动力学控制器，实现了任务空间动作到关节级别控制信号的有效转化。此外，Yao等[127]提出了由大语言模型引导的端到端框架，用于训练和部署双足机器人策略，并验证其有效性。该框架包含3个相互关联的模块：用于设计奖励函数的LLM模块；利用已有成果进行强化学习训练的模块；实现sim-to-real同态评估的模块。这些研究显著提升了双足机器人在复杂环境中的适应能力和灵活性。在以双足机器人为物理载体的具身智能体中，大模型可以调用这些预训练的动作控制模型执行特定子任务。

6.3 灵巧手控制

灵巧手控制通过精确调控机械手的多个关节和手指，实现对物体的灵活抓取、操控及操作。灵巧手通常模拟人手的结构和功能，用于完成复杂的精细操作，如抓

取、捏合和旋转物体。与机械臂相比，灵巧手的控制精度要求更高，因而其控制过程也更加复杂。早期的灵巧手操作依赖于精确编程的路径，但后来引入了基于运动和接触的解析建模。然而，这些方法在应对复杂环境和不确定性时表现一般。目前，灵巧手的控制主要通过强化学习、模型驱动学习和模仿学习等方法实现[128]。

尽管基于强化学习的数据驱动方法可以开发针对特定物体的控制策略，但在应对新物体时通常表现不佳。一些研究将强化学习与多任务学习以及适当的物体表示方法相结合，提升了灵巧手控制的泛化能力[129]。也有研究利用单个RGB摄像头，观察人类操作员的动作，以此收集远程操控的演示数据；然后，运用标准的模仿学习方法，将收集到的演示数据用于训练灵巧手的操作策略[130]。此外，一种基于渐进迁移学习的灵巧手操作框架通过重用先前训练的模型与精选样本，实现了高效的技能迁移。与传统强化学习相比，这种方法显著减少了从零开始训练所需的数据和时间，在新场景中的表现更加高效和稳定[131]。例如，DexGANGrasp[58]利用多模态大模型，实现了对环境中未知物体的精确定位，进而协助灵巧手完成抓取操作。RealDex[60]则通过将人类经验融入多模态大模型显著提升了灵巧手的抓取成功率。

总体而言，灵巧手的控制难度较高，难以直接通过具身大模型来实现端到端的动作控制。然而，随着多模态大模型的进步，以灵巧手为载体的具身智能体逐步从理论走向实际应用。

7 挑战与机遇

具身智能体领域正处于快速发展阶段，仍面临众多挑战，同时也蕴藏着新的机遇。

7.1 统一的评价体系

当前，针对不同任务的具身智能体已建立了多种评价体系，但这些体系相对零散，且各基准之间存在较大差异，难以统一且全面评估具身智能体的能力。例如：VIMABench[132]用于评估机械臂的操作与任务理解能力，但其环境感知范围未涉及三维空间；ALFWorld[133]用于评估具身智能体的任务规划能力，却缺乏对子任务操作性的评估。此外，这些评价体系大多依赖于仿真环境，而仿真器的局限性会影响评估的可靠性。因此，在更接近真实环境的场景下，亟须推出一种统一且全面的评价体系公平、准确地评估具身智能体的各项能力，包括但不限于环境感知能力、任务规划能力与动作控制能力。

7.2 高质量数据集

大模型智能表现的涌现性在很大程度上得益于海量训练数据。然而，数据集收集的高昂成本极大地限制了具身智能体领域的发展。目前，主流的做法是通过人工遥操作来收集数据，但仍然难以满足对大规模数据的需求。有些研究尝试通过具身智能体结合人工干预的方式来收集数据，但仍然无法满足训练需求[90]。为了解决这一问题，一些研究利用游戏环境（如《我的世界》）来构建具身智能体的评估体系。收集游戏世界（如高质量的游戏《黑神话：悟空》）中的海量用户行为数据，可能成为具身智能体数据集构建的可行方案。

7.3 空间环境感知

具身智能体与真实物理环境进行交互，

不仅要求具身智能体具备与视觉语言模型相似的二维视觉理解能力，还需深入理解三维空间中的位置信息与语义信息。这就要求具身智能体具备更加强大的三维感知能力，不能直接依赖现有的视觉语言模型。已有研究尝试通过点云等多模态数据增强智能体的三维感知能力，但视觉方案更符合人类的认知习惯且成本更低。为更好地提升视觉语言模型对空间信息的理解能力，有研究通过3D视觉语言数据集训练3D视觉语言模型，该模型在各种3D任务上表现出色[134]。

7.4 复杂任务规划

虽然大模型展现出较强的推理规划能力，并且提示工程、多智能体系统以及检索增强生成等方法也已被用于具身智能体的复杂任务规划，但具身智能体当前的能力仍不足以应对高难度的复杂任务，远未达到人类水平。一方面，大模型在生成结果时更多地依赖统计概率，而未能有效捕捉子任务之间的因果关系。这启示研究者在提示设计时，应注重模型对子任务之间因果联系的理解。另一方面，大模型训练数据集缺乏对复杂具身任务的描述，而人类不断接受这类具身任务的训练。例如，对于“接杯水给我”这一任务，互联网语料中很少包含对其子任务的分解。

因此，设计专门用于具身智能体与实际环境交互的大量数据集，利用此类数据集对大模型进行微调，从而增强其能力，或是从视觉输入中训练出世界模型，都是潜在的解决方案。人们可以从智能体获取的视觉信息入手，通过大量的数据学习和算法优化，让模型构建起对真实世界的认知与理解，内容涵盖物体的属性、位置关系、运动规律，以及各类环境因素等，最终训练生成能够反映真实世界特征和规律的世界模型。

7.5 全身动作控制

当前的具身智能体，尤其是由具身大模型控制的智能体，其物理载体多为自由度较少的机械臂等简单实体。然而，以人形机器人为载体的具身智能体是未来发展的必然方向，也是该领域的长远目标。现有的一种解决方案是通过分层控制，将大模型与多个预训练的动作控制模型对齐并串联。然而，这种方法存在泛化性不足和时延较高的问题。实现端到端的具身大模型控制，不仅缺乏充足的训练数据，还涉及多自由度的控制和稳定性的维持，大幅提升了任务的复杂度。因此，收集更多的全身动作控制数据，并在模型中融入更多物理规律，成为可行的解决思路。

7.6 算法效率和终端部署

随着大模型技术的发展，多种终端大模型涌现出来，如阿里的千问系列、面壁智能的终端模型、微软的Phi系列，这些模型在参数量较小的情况下也能进行终端部署。此外，一些研究在Transformer架构上进行优化，以提升具身大模型的推理速度[89]。还有研究提出了模型量化技术，以减少显存占用和推理时间。然而，这些方法通常会带来性能损失。如何在提升推理速度的同时，维持高精度并实现终端部署，依然是具身大模型落地面临的重要挑战之一。解决该问题是具身智能体大规模落地的必要前提。

7.7 反思与进化

赋予具身智能体从成功或失败案例中

自我学习、不断进化的能力，是另一个重要挑战。现有的方法大多通过提示工程或额外的记忆机制来实现这个能力，但如何从案例中有效提取知识、更新已有记忆和知识依然是一个难题。此外，增量学习也是一种解决方案，但如何平衡新旧数据、避免灾难性遗忘、优化计算资源并保证训练的稳定性，是需要深入研究的关键问题。

随着具身智能体领域近年来的快速发展，越来越多的研究者投身于这一领域。可以预见的是，这些挑战将逐步被克服，从而推动具身智能体技术的发展和应用落地。

8 结束语

具身智能体能够根据用户的指令，通过与环境的交互来完成复杂的任务。近年来，随着多模态大模型的快速发展，具身智能体也得到了长足的发展。本文对近年来具身智能体领域的一些前沿进展进行了全面、系统的综述，并对未来的发展和机遇进行了展望。具体地，本文首先介绍了多模态大模型以及具身智能体的常用数据集和常见物理载体，然后从端到端的具身大模型方案和从高级任务规划到动作控制的层级方案介绍具身智能体的实现。本文可以帮助研究者快速了解该领域的发展动向，推动具身智能体领域的发展和创新。

参考文献：

[1] GRATCH J. The promise and peril of interactive embodied agents for studying non-verbal communication: a machine learning perspective[J]. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 2023, 378(1875): 20210475.

[2] LIU Y, CHEN W X, BAI Y J, et al. Aligning cyber space with physical world: a comprehensive survey on embodied AI[J]. arXiv preprint, 2024, arXiv: 2407.06886.

[3] KOKKONEN T, EHRENBERG N, KOI P, et al. Beyond robot therapy: embodied AI, mental healthcare, and value sensitive design[M]//Social Robots in Social Institutions. Amsterdam: IOS Press, 2023.

[4] YANG W P, HU X Y, YETER I H, et al. Artificial intelligence education for young children: a case study of technology-enhanced embodied learning[J]. Journal of Computer Assisted Learning, 2024, 40(2): 465-477.

[5] TURING A M. Computing machinery and intelligence[M]//Parsing the Turing Test. Dordrecht: Springer Netherlands, 2007: 23-65.

[6] BROOKS R A. Intelligence without representation[J]. Artificial Intelligence, 1991, 47(1/2/3): 139-159.

[7] SMITH L, GASSER M. The development of embodied cognition: six lessons from babies[J]. Artificial Life, 2005, 11(1/2): 13-29.

[8] AY N, LÖHR W. The umwelt of an embodied agent: a measure-theoretic definition[J]. Theory in Biosciences, 2015, 134(3/4): 105-116.

[9] LASKEY K B. A theory of physically embodied and causally effective agency [J]. Information, 2018, 9(10): 249.

[10] BELLOT D, SIEGWART R, BESSIÈRE P, et al. Bayesian modeling and reasoning for real world robotics: basics and examples[M]//Embodied Artificial Intelligence. Heidelberg: Springer, 2004: 186-201.

[11] HAFNER V V. Agent-environment in-

teraction in visual homing[M]//Embodied Artificial Intelligence. Heidelberg: Springer, 2004: 180-185.

[12] SANDAMIRSKAYA Y, RICHTER M, SCHÖNER G. A neural-dynamic architecture for behavioral organization of an embodied agent[C]//Proceedings of the 2011 IEEE International Conference on Development and Learning (ICDL). Piscataway: IEEE Press, 2011: 1-7.

[13] SANDAMIRSKAYA Y, SCHONER G. Dynamic field theory of sequential action: a model and its implementation on an embodied agent[C]//Proceedings of the 2008 7th IEEE International Conference on Development and Learning. Piscataway: IEEE Press, 2008: 133-138.

[14] CAROLAN K, FENNELLY L, SMEATON A F. A review of multi-modal large language and vision models[EB]. arXiv preprint, 2024, arXiv: 2404.01322.

[15] WANG X, CHEN G Y, QIAN G W, et al. Large-scale multi-modal pre-trained models: a comprehensive survey[J]. Machine Intelligence Research, 2023, 20(4): 447-482.

[16] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB]. arXiv preprint, 2021, arXiv: 2103.00020.

[17] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation [EB]. arXiv preprint, 2021, arXiv: 2102.12092.

[18] BAI J, BAI S, YANG S, et al. Qwen-vl: a frontier large vision-language model with versatile abilities[EB]. arXiv preprint, 2023, arXiv: 230812966.

[19] LIU H T, LI C Y, LI Y H, et al. Improved baselines with visual instruction tuning [C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 26286-26296.

[20] SHAH D, OSINSKI B, ICHTER B, et al. LM-nav: robotic navigation with large pre-trained models of language, vision, and action[EB]. arXiv preprint, 2022, arXiv: 2207.04429.

[21] ZHOU G Z, HONG Y C, WU Q. NavGPT: explicit reasoning in vision-and-language navigation with large language models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(7): 7641-7649.

[22] YANG J N, CHEN X, QIAN S Y, et al. LLM-grounder: open-vocabulary 3D visual grounding with large language model as an agent[C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2024: 7694-7701.

[23] BROHAN A, CHEBOTAR Y, FINN C, et al. Do as I can, not as I say: grounding language in robotic affordances[EB]. arXiv preprint, 2022, arXiv: 2204.01691.

[24] LIANG J, HUANG W L, XIA F, et al. Code as policies: language model programs for embodied control[C]//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2023: 9493-9500.

[25] HUANG S Y, JIANG Z K, DONG H, et al. Instruct2Act: mapping multi-modality instructions to robotic actions with large language model[EB]. arXiv preprint, 2023, arXiv: 2305.11176.

[26] BROHAN A, BROWN N, CARBAJAL J, et al. RT-1: robotics transformer for real-world control at scale[EB]. arXiv preprint, 2022, arXiv: 2212.06817.

[27] DRIESS D, XIA F, SAJJADI M S, et al. PaLM-E: an embodied multimodal language model[EB]. arXiv preprint, 2023, arXiv: 2303.03378.

[28] WALLKÖTTER S, TULLI S, CASTELLANO G, et al. Explainable embodied agents through social cues: a review[J]. ACM Transactions on Human-Robot Interaction, 2021, 10(3): 1-24.

[29] XU Z Y, WU K, WEN J J, et al. A survey on robotics with foundation models: toward embodied AI[EB]. arXiv preprint, 2024, arXiv: 2402.02385.

[30] ALAYRAC J-B, DONAHUE J, LUC P, et al. Flamingo: a visual language model for few-shot learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-36.

[31] LI J N, LI D X, SAVARESE S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the 40th International Conference on Machine Learning. New York: ACM, 2023.

[32] HUANG S H, DONG L, WANG W H, et al. Language is not all you need: aligning perception with language models[J]. Advances in Neural Information processing Systems, 2023, 36: 72096-72109.

[33] PENG Z L, WANG W H, DONG L, et al. Kosmos-2: Grounding multimodal large language models to the world[EB]. arXiv preprint, 2023, arXiv: 2306.14824.

[34] ZHU D Y, CHEN J, SHEN X Q, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models[EB]. arXiv preprint, 2023, arXiv: 2304.10592.

[35] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning[EB]. arXiv preprint, 2023, arXiv: 2304.08485.

[36] DAI W, LI J, LI D, et al. InstructBLIP: towards general-purpose vision-language models with instruction tuning[J]. Advances in Neural Information Processing Systems, 2023, 36: 49250-49267.

[37] WANGP, BAIS, TANS N, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution [EB]. arXiv preprint, 2024, arXiv: 2409.12191.

[38] YAO Y, YU T Y, ZHANG A, et al. MiniCPM-V: a GPT-4v level MLLM on your phone[EB]. arXiv preprint, 2024, arXiv: 2408.01800.

[39] DOSOVITSKIY A. An image is worth 16×16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 201011929, 2020,

[40] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB]. arXiv preprint, 2023, arXiv: 230308774.

[41] BAI J Z, BAI S, YANG S S, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond[EB]. arXiv preprint, 2023, arXiv: 2308.12966.

[42] REID M, SAVINOV N, TEPLYASHIN D, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[EB]. arXiv preprint, 2024, arXiv: 2403.05530.

[43] CHANG A, DAI A, FUNKHOUSER T, et al. Matterport3D: learning from RGB-D data in indoor environments[C]//Proceedings of the 2017 International Conference on 3D Vision (3DV). Piscataway: IEEE Press, 2017: 667-676.

[44] DASARI S, EBERT F, TIAN S, et al. RoboNet: large-scale multi-robot learning[EB]. arXiv preprint, 2019, arXiv: 1910.11215.

[45] MAHLER J, MATL M, SATISH V, et al. Learning ambidextrous robot grasping policies[J]. Science Robotics, 2019, 4(26): eaau4984.

[46] GRAUMAN K, WESTBURY A, BYRNE E, et al. Ego4D: around the world in 3, 000 hours of egocentric video[EB].

arXiv preprint, 2022, arXiv: 2110.07058v3.

[47] SHRIDHAR M, THOMASON J, GORDON D, et al. ALFRED: a benchmark for interpreting grounded instructions for everyday tasks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10737-10746.

[48] O'NEILL A, REHMAN A, GUPTA A, et al. Open X-Embodiment: robotic learning datasets and RT-X models[EB]. arXiv preprint, 2023, arXiv: 2310.08864.

[49] WU K, HOU C K, LIU J M, et al. RoboMIND: benchmark on multi-embodiment intelligence normative data for robot manipulation[EB]. arXiv preprint, 2024, arXiv: 2412.13877.

[50] WANG Z Q, ZHENG H, NIE Y S, et al. All robots in one: a new standard and unified dataset for versatile, general-purpose embodied agents[EB]. arXiv preprint, 2024, arXiv: 2408.10899.

[51] SURATI S, HEDAOO S, ROTTI T, et al. Pick and place robotic arm: a review paper[J]. International Research Journal of Engineering and Technology, 2021, 8(2): 2121-2129.

[52] BROGÅRDH T. Present and future robot control development: an industrial perspective[J]. Annual Reviews in Control, 2007, 31(1): 69-79.

[53] MEGAHED S M. Force analysis of robot manipulators[J]. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 1989, 203(4): 217-232.

[54] LIU J M, LI C X, WANG G Q, et al. Self-corrected multimodal large language model for end-to-end robot manipulation [EB]. arXiv preprint, 2024, arXiv: 240517418.

[55] BAI Y F, LIU C K. Dexterous manipulation using both palm and fingers[C]// Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2014: 1560-1565.

[56] WAN W K, GENG H R, LIU Y, et al. UniDexGrasp: improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 3868-3879.

[57] QIN Y Z, HUANG B H, YIN Z H, et al. DexPoint: generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation[EB]. arXiv preprint, 2022, arXiv: 2211.09423.

[58] FENG Q, LEMA D S M, MALMIR M, et al. DexGANGrasp: dexterous generative adversarial grasping synthesis for task-oriented manipulation[C]//Proceedings of the 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids). Piscataway: IEEE Press, 2024: 918-925.

[59] WEI Y L, JIANG J J, XING C Y, et al. Grasp as you say: language-guided dexterous grasp generation[EB]. arXiv preprint, 2024, arXiv: 2405.19291.

[60] LIU Y M, YANG Y X, WANG Y Z, et al. RealDex: towards human-like grasping for robotic dexterous hand[EB]. arXiv preprint, 2024, arXiv: 2402.13853.

[61] CHUNG W, IAGNEMMA K. Wheeled robots[M]//Springer Handbook of Robotics. Cham: Springer, 2016: 575-594.

[62] GAO X Y, LI J H, FAN L F, et al. Review of wheeled mobile robots'navigation problems and application prospects in agriculture[J]. IEEE Access, 2018, 6: 49248-49268.

[63] LUO S Y, ZHU J, SUN P, et al. GSON: a group-based social navigation framework with large multimodal model[EB]. arXiv preprint, 2024, arXiv: 2409.18084.

[64] CUI C, MA Y S, CAO X, et al. A survey on multimodal large language models for autonomous driving[C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Piscataway: IEEE Press, 2024: 958-979.

[65] LI X H, SAVKIN A V. Networked unmanned aerial vehicles for surveillance and monitoring: a survey[J]. Future Internet, 2021, 13(7): 174.

[66] ZUO Z Y, LIU C J, HAN Q L, et al. Unmanned aerial vehicles: control methods and future challenges[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(4): 601-614.

[67] DO H T, TRUONG L H, NGUYEN M T, et al. Energy-efficient unmanned aerial vehicle (UAV) surveillance utilizing artificial intelligence(AI)[J]. Wireless Communications and Mobile Computing, 2021, 2021(1): 8615367.

[68] WU Y C, ZHANG P C, GU M Y, et al. Embodied navigation with multi-modal information: a survey from tasks to methodology[J]. Information Fusion, 2024, 112: 102532.

[69] RAIBERT M, BLANKESPOOR K, NELSON G, et al. BigDog, the rough-terrain quadruped robot[J]. IFAC Proceedings Volumes, 2008, 41(2): 10822-10825.

[70] BOUMAN A, GINTING M F, ALATUR N, et al. Autonomous spot: long-range autonomous exploration of extreme environments with legged locomotion[C]//Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2020: 2518-2525.

[71] ARM P, WAIBEL G, PREISIG J, et al. Scientific exploration of challenging planetary analog environments with a team of legged robots[J]. Science Robotics, 2023, 8(80): eade9548.

[72] NYGAARD T F, MARTIN C P, TORRESEN J, et al. Real-world embodied AI through a morphologically adaptive quadruped robot[J]. Nature Machine Intelligence, 2021, 3: 410-419.

[73] LYKOV A, LITVINOV M, KONENKOV M, et al. CognitiveDog: large multimodal model based system to translate vision and language into action of quadruped robot[C]//Proceedings of the Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. New York: ACM, 2024: 712-716.

[74] MANIATOPOULOS S, SCHILLINGER P, PONG V, et al. Reactive high-level behavior synthesis for an Atlas humanoid robot[C]//Proceedings of the 2016 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2016: 4192-4199.

[75] TANAKA F, ISSHIKI K, TAKAHASHI F, et al. Pepper learns together with children: development of an educational application[C]//Proceedings of the 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). Piscataway: IEEE Press, 2015: 270-275.

[76] XIANG J N, TAO T H, GU Y, et al. Language models meet world models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: ACM, 2023: 75392-75412.

[77] LI X H, LIU M H, ZHANG H B, et al. Vision-language foundation models as effective robot imitators[EB]. arXiv preprint, 2023, arXiv: 231101378.

[78] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: vision-language-action models transfer web knowledge to robotic control[EB]. arXiv preprint, 2023, arXiv: 2307.15818.

[79] WU H T, JING Y, CHEANG C, et al. Unleashing large-scale video generative pre-training for visual robot manipulation[EB]. arXiv preprint, 2023, arXiv: 2312.13139.

[80] CHEBOTAR Y, VUONG Q, IRPAN A, et al. Q-Transformer: scalable offline reinforcement learning via autoregressive Q-functions[EB]. arXiv preprint, 2023, arXiv: 2309.10150.

[81] BELKHALE S, DING T L, XIAO T, et al. RT-H: action hierarchies using language [EB]. arXiv preprint, 2024, arXiv: 2403.01823.

[82] MU Y, ZHANG Q L, HU M K, et al. EmbodiedGPT: vision-language pre-training via embodied chain of thought[EB]. arXiv preprint, 2023. arXiv: 2305.15021.

[83] CHEANG C L, CHEN G Z, JING Y, et al. GR-2: a generative video-language-action model with web-scale knowledge for robot manipulation[EB]. arXiv preprint, 2024, arXiv: 2410.06158.

[84] GU J Y, KIRMANI S, WOHLHART P, et al. RT-trajectory: robotic task generalization via hindsight trajectory sketches [EB]. arXiv preprint, 2023, arXiv: 2311.01977.

[85] LEAL I, CHOROMANSKI K, JAIN D, et al. SARA-RT: scaling up robotics transformers with self-adaptive robust attention[C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2024: 6920-6927.

[86] LIU J M, LIU M Z, WANG Z Y, et al. RoboMamba: multimodal state space model for efficient robot reasoning and manipulation[EB]. arXiv preprint, 2023, arXiv: 2406.04339.

[87] PEREZ E, STRUB F, DE VRIES H, et al. FiLM: visual reasoning with a general conditioning layer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 3942-3951.

[88] RYOO M S, PIERGIOVANNI A, ARNAB A, et al. TokenLearner: what can 8 learned tokens do for images and videos? [EB]. arXiv preprint, 2021, arXiv: 2106.11297.

[89] EBBESEN C L, BRECHT M. Motor cortex: to act or not to act?[J]. Nature Reviews Neuroscience, 2017, 18: 694-705.

[90] AHN M, DWIBEDI D, FINN C, et al. Autort: embodied foundation models for large scale orchestration of robotic agents[EB]. arXiv preprint, 2024, arXiv: 240112963.

[91] FIKES R E, NILSSON N J. Strips: a new approach to the application of theorem proving to problem solving[J]. Artificial Intelligence, 1971, 2(3/4): 189-208.

[92] JIANG Y Q, ZHANG S Q, KHANDELWAL P, et al. Task planning in robotics: an empirical comparison of PDDL- and ASP-based systems[J]. Frontiers of Information Technology & Electronic Engineering, 2019, 20(3): 363-373.

[93] HART P E, NILSSON N J, RAPHAEL B. A formal basis for the heuristic determination of minimum cost paths[J]. IEEE Transactions on Systems Science and Cybernetics, 1968, 4(2): 100-107.

[94] METROPOLIS N, ULAM S. The Monte Carlo method[J]. Journal of the American Statistical Association, 1949, 44 (247): 335.

[95] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]// Proceedings of the 36th International

Conference on Neural Information Processing Systems. New York: ACM, 2022: 24824-24837.

[96] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[97] SINGH I, BLUKIS V, MOUSAVIAN A, et al. ProgPrompt: generating situated robot task plans using large language models[C]//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2023: 11523-11530.

[98] HUANG W L, ABBEEL P, PATHAK D, et al. Language models as zero-shot planners: extracting actionable knowledge for embodied agents[C]//Proceedings of the 39th International Conference on Machine Learning. [S.l.]: PMLR, 2022: 9118-9147.

[99] SHINN N, CASSANO F, GOPINATH A, et al. Reflexion: Language agents with verbal reinforcement learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 8634-8652.

[100] HU Y D, LIN F Q, ZHANG T, et al. Look before you leap: unveiling the power of gpt-4v in robotic vision-language planning[EB]. arXiv preprint, 2023, arXiv: 2311.17842.

[101] ZHU M J, ZHU Y C, LI J M, et al. Language-conditioned robotic manipulation with fast and slow thinking[C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2024: 4333-4339.

[102] HUANG W L, XIA F, XIAO T, et al. Inner monologue: embodied reasoning through planning with language models [EB]. arXiv preprint, 2022, arXiv: 2207.05608.

[103] WANG G Z, XIE Y Q, JIANG Y F, et al. Voyager: an open-ended embodied agent with large language models[EB]. arXiv preprint, 2023, arXiv: 2305.16291.

[104] YANG J K, DONG Y H, LIU S, et al. Octopus: embodied vision-language programmer fromEnvironmental feedback [C]//Proceedings of the 2024 European Conference on Computer Vision. Cham: Springer, 2025: 20-38.

[105] WANG G Z, XIE Y Q, JIANG Y F, et al. Voyager: an open-ended embodied agent with large language models[EB]. arXiv preprint, 2023, arXiv: 2305.16291.

[106] MAJUMDER B P, MISHRA B D, JANSEN P, et al. Clin: a continually learning language agent for rapid task adaptation and generalization[EB]. arXiv preprint, 2023, arXiv: 2310.10134.

[107] RANA K, HAVILAND J, GARG S, et al. SayPlan: grounding large language models using 3D scene graphs for scalable task planning[C]//Proceedings of the 7th Annual Conference on Robot Learning. [S.l.]: PMLR, 2023: 23-72.

[108] GU Q, KUWAJERWALA A, MORIN S, et al. ConceptGraphs: open-vocabulary 3D scene graphs for perception and planning [C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2024: 5021-5028.

[109] WANG Z X, YU B, ZHAO J Z, et al. KARMA: augmenting embodied AI agents with long-and-short term memory systems[EB]. arXiv preprint, 2024, arXiv: 2409.14908.

[110] ZHANG Y, YANG S X, BAI C J, et al. Towards efficient LLM grounding for embodied multi-agent collaboration[EB]. arXiv preprint, 2024, arXiv: 2405.14314.

[111] KANNAN S S, VENKATESH V L N, MIN B C. SMART-LLM: smart multi-

agent robot task planning using large language models[C]//Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2024: 12140-12147.

[112] BRIENZA M, ARGENZIANO F, SURIANI V, et al. Multi-agent planning using visual language models[EB. arXiv preprint, 2024, arXiv: 2408.05478.

[113] HORI K, SUZUKI K, OGATA T. Interactively robot action planning with uncertainty analysis and active questioning by large language model[C]//Proceedings of the 2024 IEEE/SICE International Symposium on System Integration (SII). Piscataway: IEEE Press, 2024: 85-91.

[114] GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: a survey[EB]. arXiv preprint, 2023, arXiv: 2312.10997.

[115] NGUYEN H, LA H. Review of deep reinforcement learning for robot manipulation[C]//Proceedings of the 2019 Third IEEE International Conference on Robotic Computing (IRC). IEEE, 2019: 590-595.

[116] FANG B, JIA S D, GUO D, et al. Survey of imitation learning for robotic manipulation[J]. International Journal of Intelligent Robotics and Applications, 2019, 3(4): 362-369.

[117] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 3992-4003.

[118] CHERTI M, BEAUMONT R, WIGHTMAN R, et al. Reproducible scaling laws for contrastive language-image learning[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 2818-2829.

[119] YU W H, GILEADI N, FU C Y, et al. Language to rewards for robotic skill synthesis[J]. arXiv preprint, 2023, arXiv: 230608647.

[120] YE W R, ZHANG Y S, WENG H Y, et al. Reinforcement learning with foundation priors: let embodied agent efficiently learn on its own[EB]. arXiv preprint, 2024, arXiv: 2310.02635.

[121] PRATT J, DILWORTH P, PRATT G. Virtual model control of a biped walking robot[C]// Proceedings of International Conference on Robotics and Automation. Piscataway: IEEE Press, 1997: 193-198.

[122] BAO L F, HUMPHREYS J, PENG T H, et al. Deep reinforcement learning for bipedal locomotion: a brief survey[EB]. arXiv preprint, 2023, arXiv: 2404.17070.

[123] TODOROV E, EREZ T, TASSA Y. MuJoCo: a physics engine for model-based control[C]//Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. [S. l.]: PMLR, Piscataway: IEEE Press, 2012: 5026-5033.

[124] MAKOVIYCHUK V, WAWRZYNIAK L, GUO Y R, et al. Isaac gym: high performance GPU-based physics simulation for robot learning[EB]. arXiv preprint, 2021, arXiv: 2108.10470.

[125] WU F Y, GU Z Y, WU H R, et al. Infer and adapt: bipedal locomotion reward learning from demonstrations via inverse reinforcement learning[C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2024: 16243-16250.

[126] DUAN H L, DAO J, GREEN K, et al.

Learning task space actions for bipedal locomotion[C]//Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 1276-1282.

[127] YAO Y F, HE W T, GU C Y, et al. AnyBipe: an end-to-end framework for training and deploying bipedal robots guided by large language models[EB]. arXiv preprint, 2024, arXiv: 2409.08904.

[128] WEINBERG A I, SHIRIZLY A, AZULAY O, et al. Survey of learning approaches for robotic in-hand manipulation[EB]. arXiv preprint, 2024, arXiv: 2401.07915.

[129] HUANG W L, MORDATCH I, ABBEEL P, et al. Generalization in dexterous manipulation via geometry-aware multi-task learning[EB]. arXiv preprint, 2021, arXiv: 2111.03062.

[130] ARUNACHALAM S P, SILWAL S, EVANS B, et al. Dexterous imitation made easy: a learning-based framework for efficient dexterous manipulation[C]//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2023: 5954-5961.

[131] LUO Y K, LI W Y, WANG P, et al. Progressive transfer learning for dexterous in-hand manipulation with multifingered anthropomorphic hand[J]. IEEE Transactions on Cognitive and Developmental Systems, 2024, 16(6): 2019-2031.

[132] JIANG Y F, GUPTA A, ZHANG Z C, et al. VIMA: robot manipulation with multimodal prompts[C]//Proceedings of the 40th International Conference on Machine Learning. 2023: 1-48.

[133] SHRIDHAR M, YUAN X, CôTé M A, et al. Alfworld: aligning text and embodied environments for interactive learning[EB]. arXiv preprint, 2020, arXiv: 201003768.

[134] ZHU C M, WANG T, ZHANG W W, et al. LLaVA-3D: a simple yet effective pathway to empowering LMMs with 3D-awareness [EB]. arXiv preprint, 2024, arXiv: 240918125.

作者简介

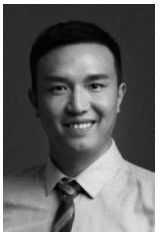
赵博涛（1997– ），男，平安科技（深圳）有限公司高级算法工程师，主要研究方向为深度学习、语音算法以及具身智能等。

亢祖衡（1989– ），男，平安科技（深圳）有限公司高级算法工程师，主要研究方向为人工智能、声纹识别、信号处理、音乐生成、大模型等。

瞿晓阳（1988– ），男，博士，平安科技前沿机器学习算法分组分责人，清华大学深圳国际研究生院校外导师，中国科学技术大学先进技术研究院校外导师，美国佛罗里达大学访问学者，主要研究方向为机器学习、大数据、体系结构、人工智能、高性能计算与存储等。

彭俊清（1973– ），男，国家认证计算机系统架构设计师，平安科技（深圳）有限公司资深经理，高级人工智能算法研究员，主要研究方向为架构设计、云平台、AI系统建设等。

张旭龙（1988– ），男，博士，平安科技（深圳）有限公司高级算法研究员，复旦大学计算机理学博士，主要研究方向为语音合成、语音转换、音频驱动虚拟人生成、音乐信息检索以及机器学习和深度学习方法在人工智能领域应用，担任清华大学深圳研究院以及中国科学技术大学先进技术研究院校外导师，目前是IEEE、中国自动化学会以及中国计算机学会会员，担任联邦数据与联邦智能专委会委员，2023年入选上海市东方英才计划青年项目。



王健宗（1983– ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理，智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会资深会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为大模型、联邦学习和深度学习等。

收稿日期: 2024-10-22

通信作者: 王健宗,jzwang@188.com

基金项目: 广东省重点领域研发计划“新一代人工智能”重大专项(No.2021B0101400003)

Foundation Item: Guangdong Province Key Field R&D Program “New Generation Artificial Intelligence” Major Special Project (No. 2021B0101400003)