大模型赋能的具身人工智能：决策与具身学习综述

WENLONG LIANG, University of Electronic Science and Technology of China, China
RUI ZHOU*, University of Electronic Science and Technology of China, China
YANG MA, University of Electronic Science and Technology of China, China
BING ZHANG, University of Electronic Science and Technology of China, China
SONGLIN LI, University of Electronic Science and Technology of China, China
YIJIA LIAO, University of Electronic Science and Technology of China, China
PING KUANG, University of Electronic Science and Technology of China, China

具身人工智能旨在开发具有物理形式的智能系统，这些系统能够在真实世界环境中进行感知、决策、行动和学习，为实现通用人工智能（AGI）提供了一种有前途的途径。尽管经过数十年的探索，具身智能体在开放动态环境中实现人类水平的通用任务智能仍然面临挑战。最近在大模型上的突破通过增强感知、交互、规划和学习革新了具身人工智能。在本文中，我们对大模型赋能的具身人工智能进行了全面调查，重点关注自主决策和具身学习。我们研究了层次化和端到端的决策范式，详细说明了大模型如何增强层次化决策的高级规划、低级执行和反馈，以及如何增强端到端决策的视觉-语言-行动（VLA）模型。对于具身学习，我们介绍了主流的学习方法，并详细说明了大模型如何深入增强模仿学习和强化学习。我们首次将世界模型整合到具身人工智能的调查中，介绍了其设计方法及在增强决策和学习中的关键作用。尽管已取得了扎实的进展，仍然存在挑战，这些挑战将在本调查的结尾处讨论，可能是进一步研究的方向。

Additional Key Words and Phrases: Embodied AI, Large Model, Hierarchical Decision-Making, End-to-end, Imitation Learning, Reinforcement Learning, World Model

1 介绍

具身人工智能 [198] 旨在开发具有实体形式的智能系统，这些系统能够在真实环境中感知、决策、行动和学习。它认为真正的智能来自于代理与其环境的互动，这为通用人工智能（AGI）提供了一个有前途的路径 [173] 。尽管对具身人工智能的探索已经跨越了几十年，但赋予代理具备人类水平的智能仍然是一个挑战，以便它们能够在开放、无结构和动态的环境中执行通用任务。早期的具身人工智能系统 [19, 189] 基于符号推理和行为主义，依赖于僵化的预编程规则，导致适应性有限和智能肤浅。尽管机器人广泛应用于制造、物流和专业操作，但它们的功能仅限于控制环境。机器学习，特别是深度学习 [123] ，的进步，标志着具身人工智能的一个转折点。视觉引导的规划和基于强化学习的控制 [162] 显著减少了代理对精确环境建模的依赖。尽管有所进展，这些模型通常是在任务特定的数据集上进行训练的，并且在泛化和可迁移性方面仍面临挑战，这限制了它们在通用应用中适应多样化场景。最近在大型模型方面的突破 [138, 139, 171, 172] ，显著提高了具身 AI 的能力。借助精准的感知、交互和规划能力，这些模型为通用具身代理 [127] 奠定了基础。然而，受大型模型驱动的具身 AI 领域仍处于初期阶段，面临着泛化、可扩展性和无缝环境交互 [166] 的挑战。迫切需要对最近在大型模型驱动具身 AI 方面的进展进行全面系统的回顾，以解决在追求 AGI 过程中存在的差距、挑战和机遇。

*Corresponding author.

Authors' Contact Information: Wenlong Liang, University of Electronic Science and Technology of China, Chengdu, China, 202421090221@std.uestc.edu.cn; Rui Zhou, University of Electronic Science and Technology of China, Chengdu, China, ruizhou@uestc.edu.cn; Yang Ma, University of Electronic Science and Technology of China, Chengdu, China, 202422090505@std.uestc.edu.cn; Bing Zhang, University of Electronic Science and Technology of China, Chengdu, China, 202521090224@std.uestc.edu.cn; Songlin Li, University of Electronic Science and Technology of China, Chengdu, China, 202422090530@std.uestc.edu.cn; Yijia Liao, University of Electronic Science and Technology of China, Chengdu, China, 202422090531@std.uestc.edu.cn; Ping Kuang, University of Electronic Science and Technology of China, Chengdu, China, kuangping@uestc.edu.cn.

通过对这些领域的全面研究，我们发现当前的研究是零散的，主题复杂但缺乏系统的分类。现有的评述主要关注于大型模型本身，如大型语言模型（LLM）[27, 140, 214] 和视觉语言模型（VLM）[94, 103, 180]，而很少关注大型模型与具身智能体的协同。尽管一些评述涉及这种整合，但它们往往专注于组件，如规划 [177]、学习 [6, 24, 193]、模拟器 [190] 和应用 [146, 190, 198]，而没有对整体范式以及这些组件如何相互作用以提高智能进行系统分析。此外，一些综合调查遗漏了最近的进展，特别是视觉-语言-行动（VLA）[107] 模型和端到端决策，这些自 2024 年以来已经变得突出。评述 [109] 提供了对 VLA 模型的详细介绍，但缺乏与分层范式的比较和对学习方法的详细探讨。此外，由于该领域的快速发展，早期的调查 [42, 209] 无法赶上最新进展。在本次调查中，我们专注于大型模型赋能的具身人工智能的决策和学习方面，分析和分类研究，找出最新进展，指出仍然存在的挑战和未来的方向，为研究人员提供一个清晰的理论框架和实际指导。我们的调查与相关调查的比较列在表格 1 中。

Table 1. **我们的调查与相关调查在调查范围上的比较。**

Survey type	Related surveys	Publication time	Large models	Decision-making		Embodied learning			World model
				Hierarchical	End2end	IL	RL	Other	
Specific	[27, 94, 103, 140, 180, 214]	2024	√	×	×	×	×	×	×
	[199]	2024	×	√	×	√	√	√	×
	[24]	2024	√	×	×	×	×	×	×
	[6, 216]	2025	×	×	×	√	√	√	×
	[177]	2024	×	√	×	×	×	×	×
	[193]	2024	×	√	×	×	×	√	×
	[154]	2025	×	×	×	×	√	×	×
	[37, 112]	2024	×	×	×	×	×	×	√
Comprehensive	[109]	2024	√	√	√	√	√	×	×
	[179]	2024	×	√	√	√	×	×	×
	[86]	2024	×	√	√	√	×	×	×
	[107]	2024	√	√	√	√	×	√	×
	Ours	–	√	√	√	√	√	√	√

我们调查的主要贡献总结如下：

(1) 从具身人工智能的角度关注大模型的赋能。在分层决策中，具身人工智能涉及高级规划、低级执行和反馈增强，因此我们根据这个层次结构回顾并分类了相关工作。在端到端决策中，具身人工智能依赖于 VLA 模型，因此我们回顾了 VLA 模型及其增强。由于主要的具身学习方法是模仿学习（IL）和强化学习（RL），我们回顾了大模型如何在模仿学习中赋能策略和战略网络构建，以及在强化学习中如何赋能奖赏函数设计和策略网络构建。

(2) 全面回顾具身决策和具身学习。在本综述中，我们对由大型模型赋能的具身人工智能的决策和学习进行了全面回顾。对于决策，我们回顾了由大型模型赋能的分层和端到端范式，并详细比较了它们。对于具身学习，我们不仅回顾了模仿学习和强化学习，还包括迁移学习和元学习。此外，我们还回顾了世界模型及其如何促进决策和学习。

(3) 双重分析方法以深入洞察。我们采用一种双重分析方法论，结合横向和纵向视角。横向分析审查并比较多种方法，例如多种大型模型、分层与端到端决策、模仿学习与强化学习，以及多种具体学习策略。纵向分析追踪核心模型或方法的演变，详细描述它们的起源、后续进展和开放挑战。这种双重方法论使得主流方法的具体 AI 在宏观层面概览和深入洞察成为可能。

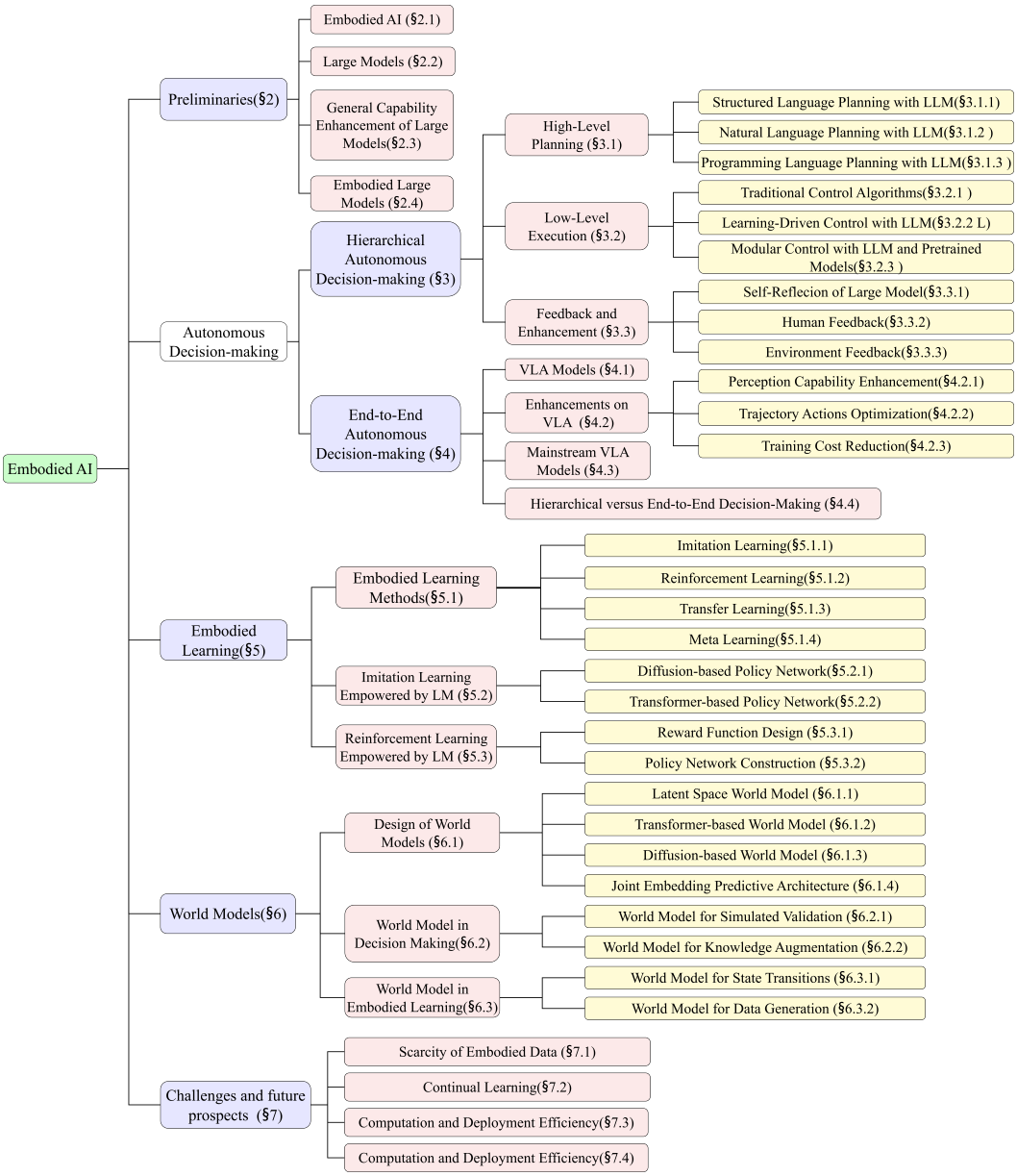
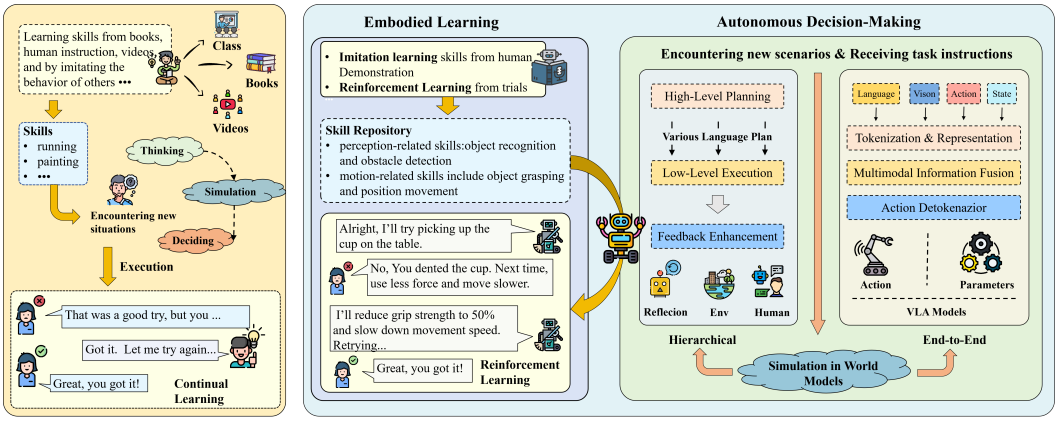


Fig. 1. 本调查的组织结构。

我们的调研组织如图 1 所示。第 2 节介绍了具身人工智能的概念，概述了大型模型，并讨论了它们的通用能力提升。然后，它展示了大型模型在具身人工智能上的协同效应，为后续部分奠定基础。第 3 节深入探讨了分层决策范式，详细说明了大型模型如何通过反馈赋能动态的高层次规划、低层次执行和迭代优化。第 4 节聚焦于端到端决策，首先介绍并分解了 VLA 模型，然后探讨了在感知、动作生成和部署效率方面的最新增强。在本节结束时，提供了分层和端到端决策之间的综合比较。第 5 节介绍了具身学习方法，特别是大型模型



(a) Human paradigm (b) Robot paradigm

Fig. 2. **具身人工智能：从整个过程所需能力的前景来看。**

增强的模仿学习和强化学习。第 6 节介绍了世界模型并讨论了它们在决策和具身智能学习中的作用。第 7 节讨论了开放挑战并指出未来展望。第 8 节对调研进行了总结。

2 预备知识

大型模型 [27, 140, 214] 在近年来展示了令人印象深刻的能力，并获得了极大的关注。研究人员已经开始利用这些模型来构建人工智能代理 [3, 76, 127, 222] 。在本节中，我们提供关于具身人工智能和大型模型的基础知识。我们首先介绍具身人工智能的基本概念和整体过程。随后，我们展示主流的大型模型及其增强通用能力的技术。最后，我们讨论大型模型在具身人工智能系统中的应用。

2.1 具身人工智能

一个具身 AI 系统通常由两个主要组件组成：物理实体和智能代理 [198] 。物理实体，如人形机器人 [114] 、四足机器人 [11] 和智能车辆 [149] ，执行动作并接收反馈，作为物理世界和数字世界之间的接口。智能代理构成认知核心，使自主决策和学习成为可能。为了执行具身任务，具身 AI 系统从语言指令中解释人类意图，主动探索周围环境，感知环境中的多模态元素，并执行任务的动作。这个过程模仿了人类学习和解决问题的范式。如图 2 （a）所示，人类从各种资源中学习技能，例如书籍、教学材料和在线内容。当遇到不熟悉的场景时，他们评估环境，计划必要的行动，进行策略的心理模拟，并根据结果和外部反馈进行自我调整。具身代理模拟这种类人学习和解决问题的范式，如图 2 （b）所示。通过模仿学习，代理从人类演示或视频数据中获得技能。当面对复杂任务或新环境时，他们分析周围环境，根据目标分解任务，自主制定执行策略，并通过模拟器或世界模型优化计划。执行后，通过整合外部反馈，强化学习优化策略和行动，提高整体性能。

具身智能的核心是使代理能够在开放动态环境中自主地做出决策和学习新知识 [216] 。自主决策可以通过两种方法实现：（1）分层范式 [3] ，该范式将感知、规划和执行分为不同的模块，（2）端到端范式 [222] ，该范式将这些功能集成到一个统一的框架中，实现无缝操作。具身学习使代理能够通过长期的环境交互自主优化其行为策略和认知模型，实现持续改进。它可以通过模仿学习 [18] 从示范中获取技能，并通过强化学习 [9] 在任务执行过程中通过迭代完善来优化技能。此外，世界模型 [221] 也在为代理提供机会进行尝试和积累经验方面起着关键作用，通过模拟现实世界的推理空间。这些组件协同工作以增强具身代理的能力，向 AGI 迈进。

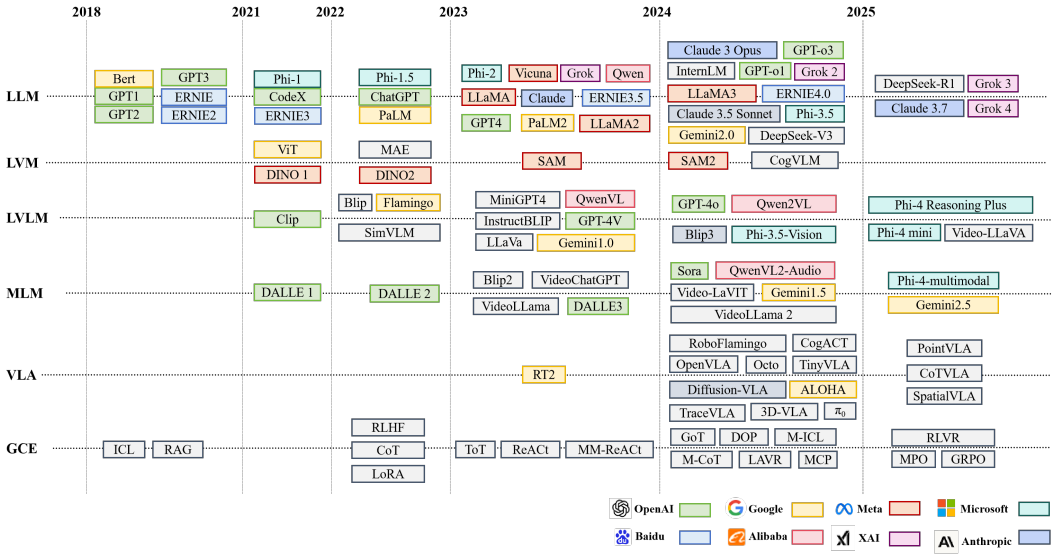


Fig. 3. **主要的大型模型时间线。**

2.2 大型模型

大模型，包括大型语言模型（LLM）、大型视觉模型（LVM）、大型视觉-语言模型（LVLM）、多模态大型模型（MLM）以及视觉-语言-动作（VLA）模型，在架构、数据规模和任务复杂性方面取得了显著突破，展现了强大的感知、推理和交互能力。图 3 显示了主要大模型的时间线以及对它们的一般能力增强（GCE）。

在 2018 年，谷歌发布了 BERT，一种通过自监督任务预训练的双向 Transformer 模型，显著提高了自然语言任务的表现。随后，OpenAI 发布了 GPT，这是一种基于 Transformer 架构的生成模型，使用大规模无监督语料库的自回归训练来生成连贯的文本，标志着生成模型的突破。GPT-2 进一步扩大了模型的规模和训练数据，提高了文本的连贯性和自然性。在 2020 年，GPT-3 以其庞大的模型容量和多样的训练数据设立了一个里程碑，在文本生成、问答和翻译方面表现出色。它首次展示了零样本和少量样本学习的能力，为未来的研究铺平了道路。之后，Codex 通过代码数据集的预训练，推进了代码生成和理解。ChatGPT（基于 GPT-3.5）使得与用户的互动更加自然和流畅，同时支持广泛的知识领域。谷歌的 PaLM 通过大规模训练和优化计算在语言理解、生成和推理方面表现优异。InstructGPT 建立于 GPT-3 之上，利用人类反馈的强化学习（RLHF）来符合人类偏好。Meta 的 Vicuna，一个开源对话模型，以低计算成本提供高质量的交互，非常适合资源受限的系统。Meta 的 Llama 系列（7B、13B、30B、65B 参数）对开源研究和开发做出了显著贡献。

2.2.1 **大型视觉模型** . LVM [87] 的功能是处理视觉信息。Vision Transformer (ViT) [39] 适配了 Transformer 架构用于计算机视觉，将图像划分为固定大小的块，并使用自注意力机制捕获全局依赖。在此基础上，Facebook AI 发布了 DINO [25] 和 DINOv2 [126] ，利用 ViT 进行自监督学习。DINO 采用一种自蒸馏的方法，利用师生网络生成高质量的图像表征，通过自注意力和对比学习，在没有标注数据的情况下捕获语义结构。DINOv2 通过改进的对比学习和更大的训练集增强了 DINO，提高了表征质量。Masked Autoencoder (MAE) [70] 利用自监督学习来重建被遮盖的视觉输入，从而能够对大量未标记的图像数据集进行预训练。Segment Anything Model (SAM) [88, 145] 在 1100 万张图像上进行预训练，支持多种分割任务，包括语义分割、实例分割和对象分割，并通过基于用户反馈的微调展现出强大的适应性。

2.2.2 **大型视觉语言模型**. LVLM [97] 集成了预训练视觉编码器和视觉语言融合模块，可以通过语言提示处理视觉输入并对视觉相关的查询做出响应。由 OpenAI 开发的 CLIP [137]，利用对比学习 [32] 在大规模图文对上训练图像和文本编码器，使配对样本特征对齐，同时最小化未配对样本，以创建匹配文本语义的视觉表示。BLIP [98] 采用双向自监督学习来融合视觉和语言数据，使用“引导”策略来提升预训练效率，并改进视觉问答和图像字幕的性能。BLIP-2 [97] 进一步引入了 QFormer 结构，从冻结的图像编码器中提取视觉特征，并通过多模态预训练将它们与语言指令对齐，以实现高效的跨模态融合。Flamingo [5] 在小样本学习中表现出色，处理多模态数据时使用最少的样本以支持数据稀缺场景中的跨模态推理。GPT-4V [212] 扩展了传统的 GPT，以处理联合图文输入，生成图像描述并通过稳健的多模态推理回答视觉问题。DeepSeek-V3 [105] 通过采用动态稀疏激活架构进一步扩展了多模态推理的边界。它引入了一种结合任务特定专家和动态参数分配的混合路由机制，在跨模态融合任务中实现了高计算效率。

2.2.3 **多模态大型模型**. MLM 可以处理多种模态，包括文本、视觉、音频等。根据输入输出范式，MLM 可以分为多模态输入文本输出模型和多模态输入多模态输出模型。

多模态输入文本输出模型整合了多样化的数据模态，以实现全面的内容理解。例如，Video-Chat [99] 通过会话建模增强了视频分析，在动态视觉内容理解方面表现出色。基于 Llama 架构，VideoLLaMA [207] 融合了视觉和音频输入，实现了稳健的视频内容分析。谷歌的 Gemini [168] 专为多模态设计，能够高效处理文本、图像和音频，用于图像描述和多模态问答。PaLM-E [40] 将多模态输入转换为统一的向量，并将其输入到 PaLM 模型进行端到端训练，实现了强大的多模态理解。

多模态输入多模态输出模型通过学习复杂的数据分布来生成多种数据模态，例如文本、图像和视频。例如，DALL · E [144]，通过扩展带有向量量化变分自编码器（VQ-VAE）和一个 12 亿参数的 Transformer 的 GPT-3，生成了创造性和与提示对齐的图像，支持零样本任务。DALL · E2 [143] 通过将 CLIP 整合到其中，采用了一个两阶段的过程：首先生成低分辨率图像，然后进行超分辨率增强，从而极大地改善了图像的质量和多样性。DALL · E3 [14] 通过增强文本编码器和提高训练数据质量进一步优化了图像-提示的对齐。2024 年，OpenAI 发布了 Sora [20]，这是一种视频生成模型，可以从文本提示生成长达 60 秒的高质量连贯视频。Sora 利用编码网络将输入转换为离散标记，利用大规模扩散模型优化序列，并将去噪后的标记投影回视频空间。

2.2.4 **视觉-语言-动作模型**. VLA 模型最近获得了极大的关注。其核心目标是直接将多模态输入映射到动作输出，而不是通过分层决策的中间步骤，从而提高机器人的感知-动作集成能力。VLA 的概念首次由 RT-2 [222] 提出，该模型利用预训练的视觉-语言模型将动作空间离散化为动作标记，并通过互联网数据和机器人数据的联合微调实现了泛化。然而，其离散的动作设计和闭源性质限制了其灵活性和进一步研究。为了克服这些限制，基于连续动作生成的 VLA 模型 [101] 和开源 VLA 模型 [84] 出现了。最近对 VLA 模型的研究进一步解决了这些挑战。BYO-VLA [66] 、3D-VLA [215] 、PointVLA [95] 处理了视觉输入处理。Octo [169] 和 Diffusion-VLA [185] 解决了动作生成的准确性问题。TinyVLA [187] 和 π_0 [16] 改善了计算效率。

2.3 大模型通用能力提升

大型模型仍然存在推理能力、幻觉、计算成本和任务特异性上的限制。研究人员提出了一系列技术来提高它们的总体能力，正如图 4 中所示。

上下文学习 (ICL) [21] 使大型模型通过精心设计的提示实现零样本泛化，允许它们在无需额外训练和调优的情况下解决新任务。利用输入提示中的上下文，大型模型可以理解任务要求并生成相关的输出，使其成为一个从自然语言处理到特定任务问题解决等应用中的多功能工具。最近的进展集中在优化提示技术，如自动提示生成和动态实例选择，以增强 ICL 在不同领域的稳健性。

思想之链（XoT）是一系列推理框架，旨在提升大型模型解决数学、逻辑和开放性问题的能力。思维链（CoT）[184] 在提示中加入中间推理步骤，引导大型模型将复杂问题分解为易于处理的部分。思维树（ToT）[202] 通过在类似树的结构中探索多条推理路径来扩展 CoT，使大型模型能够评估替代解决方案并在必要时回溯。思维图（GoT）[13] 通过采用图结构进一步拓展 ToT，其中节点表示中间状态，边捕捉关系和依赖性，支持灵活的非线性推理。

增强检索生成（RAG）[93] 从外部知识库中检索相关信息，如数据库和网络资源，并将其提供给大型模型以获得准确的响应。RAG 缓解了大型模型知识过时或不完整的问题，确保可以访问最新的和特定领域的信息。最近的进展包括结合稠密和稀疏检索方法的混合检索机制，以平衡精度和效率，以及微调策略以有效地将检索到的内容与生成的输出对齐。

推理与行动（ReAct）[203] 结合了推理与行动执行，能够在完成任务的过程中产生明确的推理轨迹。通过要求大型模型在行动之前表达他们的思维过程，ReAct 提高了决策透明度，并改善了动态互动环境中的性能。

从人类反馈中进行强化学习（RLHF）[128] 将人类偏好整合到大型模型的训练中，使大型模型与人类的价值观和意图一致。利用人类反馈作为奖励信号，RLHF 改善了模型在动态、互动设置中生成有用、无害和诚实输出的能力。通过提示模型生成多个响应，RLHF 允许人类根据质量和安全性对它们进行排名或评分，并使用这些反馈来优化模型的未来生成，确保一致性和伦理考虑。

模型上下文协议（MCP）[73] 是由 Anthropic 引入的一个开源标准，它为大型模型与外部数据源、工具和服务的交互提供了标准化接口。MCP 增强了大型模型的互操作性和适应性，使其能够与不同外部系统无缝集成。MCP 的最新发展集中在扩展其与多模态输入的兼容性和优化其在实时应用中的性能。

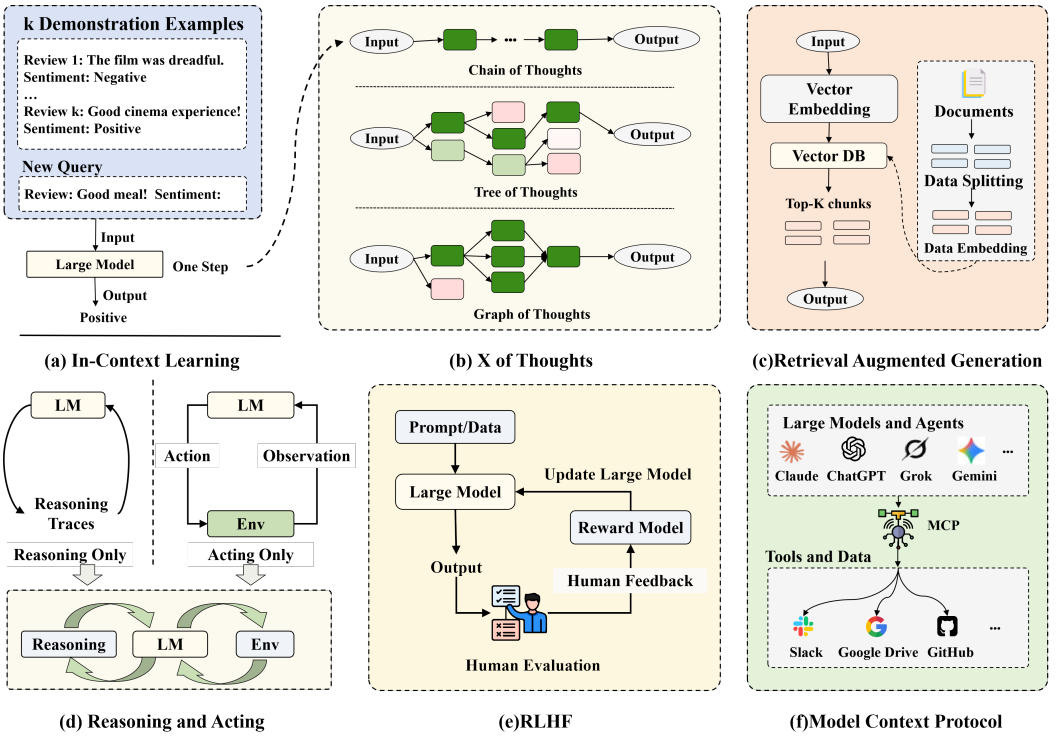


Fig. 4. 大型模型的通用能力增强。

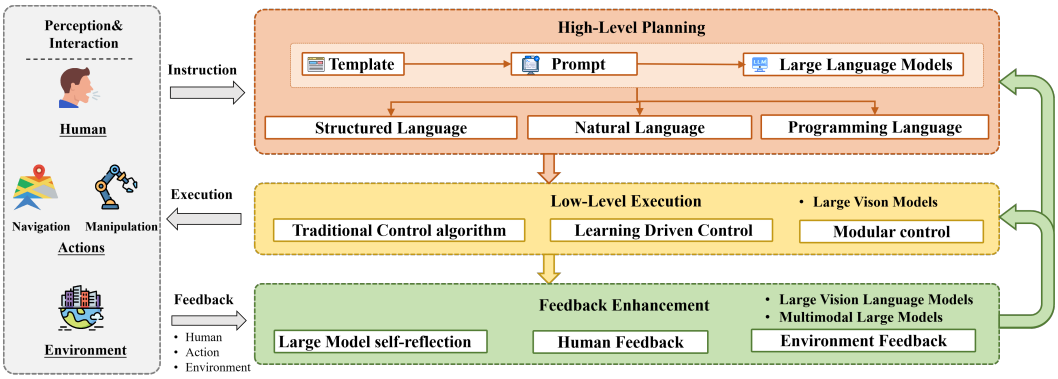


Fig. 5. 层次决策范式，包括感知和交互、高层规划、低层执行、反馈和增强。

2.4 具身大模型

大型模型通过增强智能代理的能力来赋能具身智能。通过包括文本、视觉、音频和触觉在内的多种模态的无缝集成，具身大型模型（ELM），也称为具身多模态大型模型（EMLM），可以赋能代理构建能够在复杂环境中感知、推理和行动的复杂系统，在自主决策和具身学习中起到关键作用。

不同的大模型为具身智能体赋予了不同的能力。LLM 通常作为认知骨干，用于处理自然语言输入，理解上下文细微差别，并生成可操作的响应。LVM 通常用于感知任务或在任务执行期间作为可调用的 API，利用预训练的视觉编码器来预测物体类别、姿态和几何形状。LVLM 和 MLM 可以通过与多种模式整合 LLM 进一步增强智能体的能力，使代理能够理解跨文本、视觉和音频的人类指令，生成与上下文相关的响应或行为。近期在复杂导航和操作任务中的进展突出了 MLM 的优势。与以前单独处理功能的模型不同，VLA 模型从视觉和语言输入到可执行动作学习端到端映射。这个简化的流程使智能体能够解释复杂的指令，感知动态环境，并执行精确的物理动作，从而形成更强大和多功能的具身 AI 系统。除了增强计划智能之外，研究人员正越来越多地探索其生成能力以推进具身学习并协助构建世界模型，从而进一步支持通向 AGI 的道路。

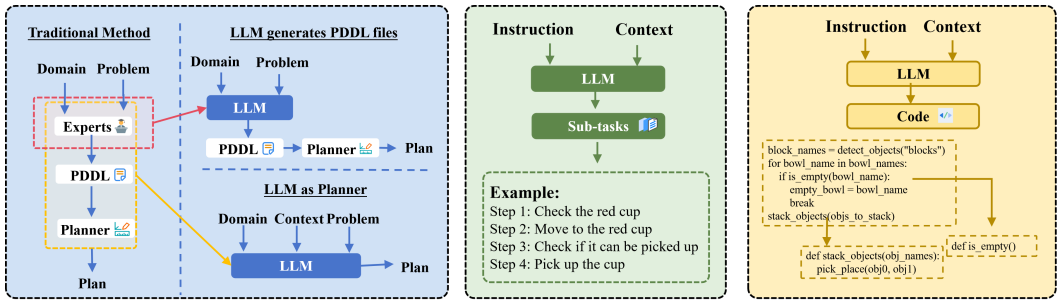
自主代理的决策旨在将环境感知和任务理解转化为可执行的决策和物理动作。传统的决策采用分层范式，包括感知与交互、高层规划、低层执行，以及反馈与增强。感知与交互层依赖于视觉模型，高层规划层依赖于预定义的逻辑规则 [47]，而低层执行层依赖于经典控制算法 [48]。这些方法在结构化环境中表现出色，但在非结构化或动态环境中却表现不佳，原因在于整体优化和高层决策的局限性。

大型模型在其稳健的学习、推理和泛化能力方面的进展，显示了在复杂任务处理中的前景。通过将大型模型的推理能力与物理实体的执行能力相结合，这为自主决策提供了一种新的模式。如图 5 所示，环境感知首先诠释了代理的周围环境，借助 LLM 的大规模模型，高层次规划随后基于感知信息和任务指令将复杂任务分解为子任务，借助 LLM 的大规模模型，低层次执行然后将子任务转换为精确的物理动作，最后，借助 LLM 的大规模模型的反馈增强引入闭环反馈以增强智能。

2.5 高级规划

高层规划根据任务指令和感知信息生成合理的计划。传统的高层规划依赖于基于规则的方法 [51, 67, 116]。给定在规划域定义语言（PDDL）中指定的初始状态和目标，启发式搜索规划器验证动作的前提条件以确定可行性，并采用搜索树选择最优的动作序列，从而生成一个高效且经济的计划 [81]。尽管在结构化环境中有效，基于规则的方法在非结构化或动态

场景中适应性较差。大型模型利用其零样本和少样本泛化能力，推动了在应对这些挑战方面的突破。根据规划的形式，赋能于大型语言模型的高层规划可以被分类为结构化语言规划、自然语言规划和编程语言规划，如图 6 所示。



(a)Structured Language Planning (b)Natural Language Planning (c)Programming Language Planning

Fig. 6. **由大型模型赋能的高级规划。**

2.5.1 **使用 *LLM* 的结构化语言规划**. LLM 可以通过两种关键策略来增强结构化语言规划，如图 6 (a) 所示。（1）第一种策略是将 LLM 用作规划者，利用其零/小样本泛化能力来生成计划。然而，Valmeekam 等人 [174] 证明由于严格的 PDDL 语法和语义，LLM 经常生成不可行的计划，导致逻辑错误。为缓解这一问题，LLV [7] 引入了一个外部验证器，即 PDDL 解析器或环境模拟器，通过错误反馈来检查和迭代改进 LLM 生成的计划。FSP-LLM [164] 优化了提示工程，以使计划符合逻辑约束，确保任务的可行性。（2）第二种策略利用 LLM 自动生成 PDDL，减少领域建模中的手动努力。在 LLM+P [106] 中，LLM 创建了 PDDL 领域文件和问题描述，随后由传统规划器解决，将语言理解与符号推理相结合。PDDL-WM [56] 使用 LLM 迭代构建和改进 PDDL 领域模型，这些模型通过解析器和用户反馈进行验证，以确保正确性和可执行性。通过将 LLM 用作直接规划者或 PDDL 生成器，这些策略提高了自动化程度并减少了用户参与，从而显著提高了规划效率、适应性和可扩展性。

2.5.2 **使用 *LLM* 进行自然语言规划**. 自然语言提供比结构化语言更大的表达灵活性，使得能够充分利用 LLM 将复杂计划分解为子计划 [100, 156]，如图 6 (b) 所示。然而，自然语言规划通常会产生不可行的计划，因为其输出往往基于经验而非实际环境。例如，当被要求“清洁房间”时，LLM 可能会建议“取来吸尘器”，而不去验证其可用性。Zero-shot [76] 探索了使用 LLM 将高级任务分解为一系列可执行的语言规划步骤的可行性。他们的实验表明，LLM 可以基于常识推理生成初步计划，但缺乏对物理环境和行动可行性的约束。

为了解决这个问题，SayCAN [3] 将 LLM 与强化学习结合起来，将 LLM 生成的计划与预定义的技能库和价值函数结合起来，以评估行动的可行性。通过为动作评分期望的累积奖励，SayCAN 过滤掉不切实际的步骤（例如，“跳上桌子去抓杯子”），选择更安全的高价值动作（例如，“走到桌子旁并伸出手”）。Text2Motion [104] 通过结合几何可行性进一步增强涉及空间交互的任务规划。它使用 LLM 提出候选动作序列，然后由检测器评估其物理可行性，以避免执行诸如“将大箱子堆放在小球上”的动作。然而，这两种方法都依赖于固定的技能集，缺乏适应开放式任务的能力。Grounded Decoding [78] 通过引入灵活的解码策略解决了这一限制。它动态地将 LLM 输出与实时的基础模型结合，该模型根据当前的环境状态和代理能力评估动作可行性，引导 LLM 生成上下文上可行的计划。

2.5.3 **使用大型语言模型进行编程语言规划**. 编程语言规划将自然语言指令转化为可执行程序，利用代码的精确性来定义空间关系、函数调用和控制 API，以便在如图 6 (c) 所示的实体任务中进行动态的高层次规划。CaP [102] 将任务规划转化为代码生成，产生具有递归定义

函数的 Python 风格程序，以创建动态函数库。例如，在机器人导航中，CaP 首先定义一个“移动”函数，然后根据任务需求扩展为“避障移动”或“靠近目标”。这种自我扩展的库增强了对新任务的适应性，而无需预定义模板。然而，CaP 对感知 API 的依赖和不受限制的代码生成限制了其处理复杂指令的能力。为克服这些限制，Instruct2Act [75] 利用多模态基础模型提供了更集成的解决方案，以统一感知、规划和控制。它使用视觉语言模型进行准确的对象识别和空间关系理解，以提供精确的环境感知。然后将感知数据传递给 LLM，从预定义的机器人技能库生成基于代码的动作序列。该方法显著提高了规划准确性，使代理能够有效适应新环境，特别是在具有显著视觉组件的任务中。ProgPrompt [165] 采用结构化提示，包括环境操作、对象描述和示例程序，引导 LLM 生成定制的、基于代码的计划。通过加入预定义的约束，ProgPrompt 最小化了无效代码生成，并增强了跨环境的适应性。

继高层任务规划之后，低层动作使用预定义的技能列表 [76] 来执行。技能列表代表了体现化代理执行特定任务所需的一系列基本能力或动作模块。它们是任务规划与物理执行之间的桥梁。例如，与感知相关的技能包括物体识别和障碍检测，而与运动相关的技能包括抓取和移动物体。低层技能的实现涉及控制理论、机器学习和机器人工程。方法从传统的控制算法发展到以学习为驱动的控制再到模块化控制，如图 7 所示。

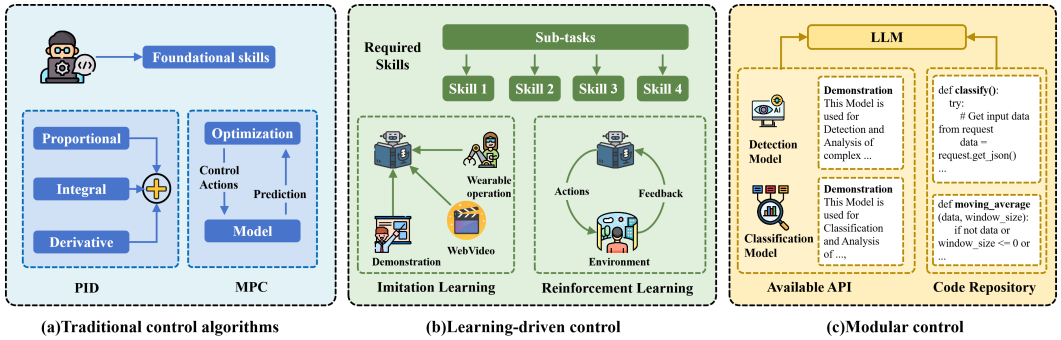


Fig. 7. **低级执行**。

2.5.4 **传统控制算法**. 基础技能通常使用传统控制算法设计，这些算法主要利用具有清晰数学推导和物理原理的经典基于模型的技术。比例-积分-微分（PID）控制 [73] 调整参数以最小化机器人手臂关节控制中的误差。状态反馈控制 [167] 常与线性二次调节器（LQR）[115] 配合使用，通过系统状态数据优化性能。模型预测控制（MPC）[1] 通过滚动优化预测状态并生成控制序列，非常适合无人机路径跟踪等任务。传统控制算法提供了数学可解释性、低计算复杂度和实时性能，从而实现可靠的任务执行。然而，当面对动态环境时，传统控制算法缺乏适应性，难以处理高维不确定系统动力学。它们需要与数据驱动技术集成，例如深度学习和强化学习，以增强泛化能力。例如，当四足机器人在不平坦的地形上导航时，传统的 PID 控制与学习算法协作，以动态调整其步态。

2.5.5 **基于学习的 *LLM* 控制**. 机器人学习位于机器学习和机器人技术的交汇处。它使智能体能够从广泛的数据中开发控制策略和低级技能，包括人类演示、模拟和环境交互。模仿学习和强化学习是实现这一目标的两个重要学习方法。模仿学习从专家演示中训练策略，减少探索时间并实现快速策略开发。Embodied-GPT [121] 利用一个 7B 语言模型进行高层规划，并通过模仿学习将计划转化为低级策略。强化学习通过迭代试验和环境奖励优化策略，适用于高维动态环境。Hi-Core [130] 采用一个两层框架，其中 LLM 设定高层策略和子目标，而强化学习在低层生成具体动作。这些由 LLM 驱动的学习控制方法提供了强大的适应性和

泛化能力。然而，它们的训练通常需要大量的数据和计算资源，策略的收敛性和稳定性难以保证。

2.5.6 **使用 *LLM* 和预训练模型的模块化控制**. 模块化控制将 LLM 与预训练的策略模型（如用于视觉识别的 CLIP [137] 和用于分割的 SAM [87] ）整合。在为 LLM 提供这些工具的描述后，它们可以在任务执行过程中动态调用。DEPS [181] 结合多个不同的模块，根据任务需求和预训练模型的自然语言描述完成检测和操作。PaLM-E [40] 将 LLM 与用于分割和识别的视觉模块融合。CLIPort [161] 利用 CLIP 进行开放词汇检测。[102] 利用 LLM 生成代码来创建可调用函数库以实现导航和操作。通过利用共享的预训练模型，这种模块化方法确保了在各种任务中的可扩展性和可重用性。

然而，存在一些挑战。首先，调用外部策略模型可能会引入额外的计算和通信延迟，特别是在实时任务中（例如，自动驾驶 [205] ），这类延迟可能会显著影响响应效率。其次，智能体的整体性能高度依赖于预训练的策略模型的质量。如果策略模型存在缺陷（如泛化能力不足或训练数据偏差），即使拥有强大的 LLM 规划能力，其执行结果仍可能令人不满意。因此，优化模块间的通信效率，提升策略模型的鲁棒性，以及设计更智能的调用决策机制是非常重要的。

2.6 反馈与提升

分层决策架构通过任务描述和示例提示指导任务规划。为了确保任务规划的质量，应引入闭环反馈机制。反馈可能来自大模型本身、人类和外部环境，如图 8 所示。

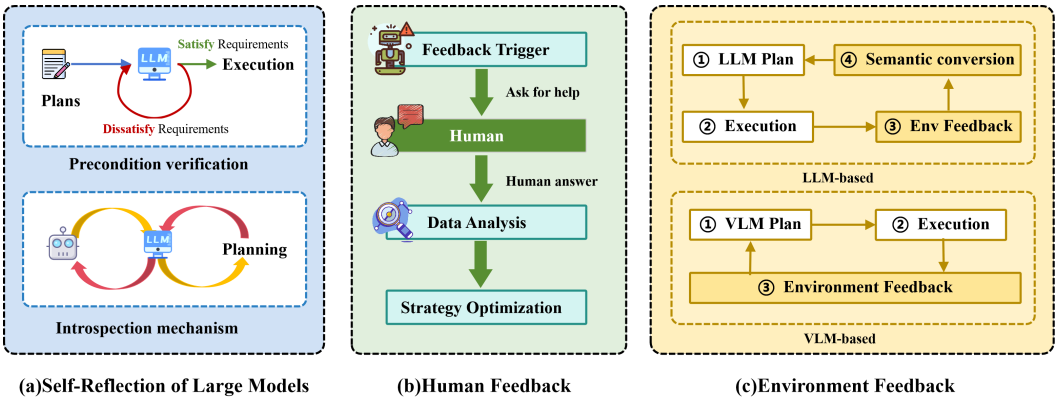


Fig. 8. **反馈和增强**。

2.6.1 **大模型的自我反思**. 大模型可以作为任务规划者、评估者和优化者，从而在没有外部干预的情况下迭代优化决策过程。代理得到行动反馈，自主检测和分析失败的执行，并不断从以前的任务中学习。通过这种自我反思和优化机制，大模型可以生成稳健的策略，在长序列规划、多模态任务和实时场景中提供优势。自我反思可以通过两种方式实现，如图 8 (a) 所示。

(1) 第一种方法通过重新提示 [142] 来触发计划再生，基于检测到的执行失败或前置条件错误。重新提示将错误上下文（例如，在开门前未能解锁门）作为反馈来动态调整提示，从而纠正 LLM 生成的计划。DEPS [142] 采用“描述、解释、计划、选择”框架，其中 LLM 描述执行过程，解释失败原因，并重新提示以纠正计划缺陷，增强交互式计划。

(2) 第二种方法采用自省机制，使大型语言模型能够独立评估和优化其输出。Self-Refine [111] 使用单一的大型语言模型进行规划和优化，通过多次自我反馈周期迭代地提高计划的

合理性。Reflexion [159] 通过结合长期记忆来存储评估结果，结合多种反馈机制来增强计划的可行性，从而对其进行扩展。ISR-LLM [220] 在基于 PDDL 的规划中应用迭代自我优化，生成初步计划，执行合理性检查，并通过自我反馈优化结果。Voyager [178] 针对编程语言规划进行定制，通过从执行失败中提取反馈构建动态代码技能库，使代理能够适应复杂任务。

2.6.2 **人类反馈** . 人类反馈通过与人类建立交互闭环机制来增强计划的准确性和效率，如图 8 (b) 所示。这种方法使代理能够根据人类反馈动态调整行为。KNOWNO [150] 引入了一种不确定性测量框架，使大语言模型（LLM）能够识别知识空白，并在高风险或不确定的场景中寻求人类帮助。EmbodiedGPT [122] 采取了一个计划-执行-反馈循环，当低层控制失败时，代理请求人类输入。这种人类反馈结合强化学习和自监督优化，使代理能够迭代地改进其计划策略，确保更好地适应动态环境条件。YAY Robot [157] 允许用户通过命令暂停机器人并提供指导，促进实时的基于语言的纠错。反馈被记录用于策略微调和定期查询，实现实时和长期的改进。IRAP [72] 允许与人类进行交互式问答，以获取任务特定的知识，实现精确的机器人指令。

2.6.3 **环境反馈** . 环境反馈通过环境的动态交互增强了基于 LLM 的规划，如图 8 (c) 所示。Inner Monologue [79] 将多模态输入转换为语言描述进行“内心独白”推理，使 LLM 能够根据环境反馈调整计划。TaPA [192] 整合了开放词汇对象检测，并针对导航和操作量身定制计划。DoReMi [57] 检测计划与实际结果之间的差异，并采用多模态反馈动态调整任务。在多代理设置中，RoCo [113] 利用环境反馈和代理间的通信实时纠正机械臂路径规划。

基于大型语言模型的规划通常需要将反馈转换为自然语言。视觉语言模型通过整合视觉输入和语言推理简化了这一过程，避免了反馈转换。ViLaIn [160] 将大型语言模型与视觉语言模型结合起来，从语言指令和场景观察生成机器可读的 PDDL，以高精度驱动符号规划器。ViLa [74] 和 Octopus [200] 通过利用 GPT4-V 的多模态语言模型生成计划，整合感知数据以实现稳健的零样本推理，从而实现机器人视觉语言规划。Voxposer [77] 进一步利用多模态语言模型提取空间几何信息，从机器人观察中生成三维坐标和约束图，以填充代码参数，从而在规划中提高空间准确性。

3 端到端自主决策

层级范式依赖于单独的任务规划、行动执行和反馈模块，因此它容易出现错误累积，并且在不同任务间难以泛化。此外，从大型模型中提取的高层语义知识难以直接应用于机器人行动执行，导致整合差距。为了缓解这些挑战，端到端自主决策最近受到极大关注，这种方法直接将多模态输入（即视觉观察和语言指令）映射到行动。它通常通过 VLA 实现，如图 9 所示。

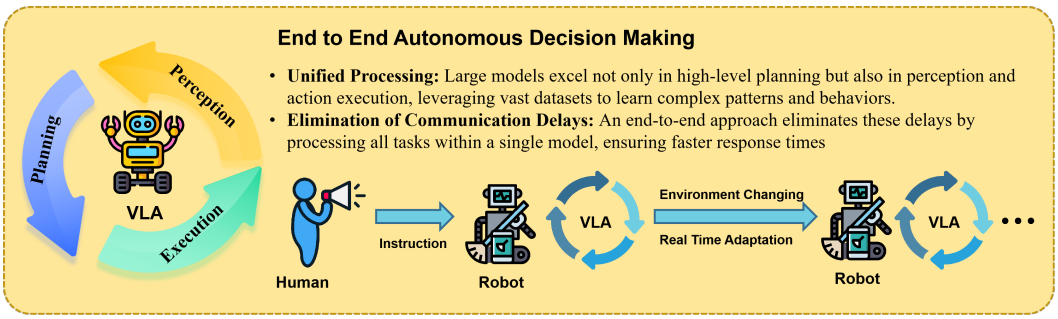


Fig. 9. 由 VLA 进行端到端决策。

3.1 视觉-语言-动作模型

VLA 模型通过将感知、语言理解、计划、动作执行和反馈优化整合到一个统一的框架中，代表了具身人工智能的突破。通过利用大型模型的丰富先验知识，VLA 模型能够在动态、开放的环境中实现精确且适应性强的任务执行。一个典型的 VLA 模型由三个关键组件组成：标记化和表示、多模态信息融合以及动作去标记化，如图 10 所示。

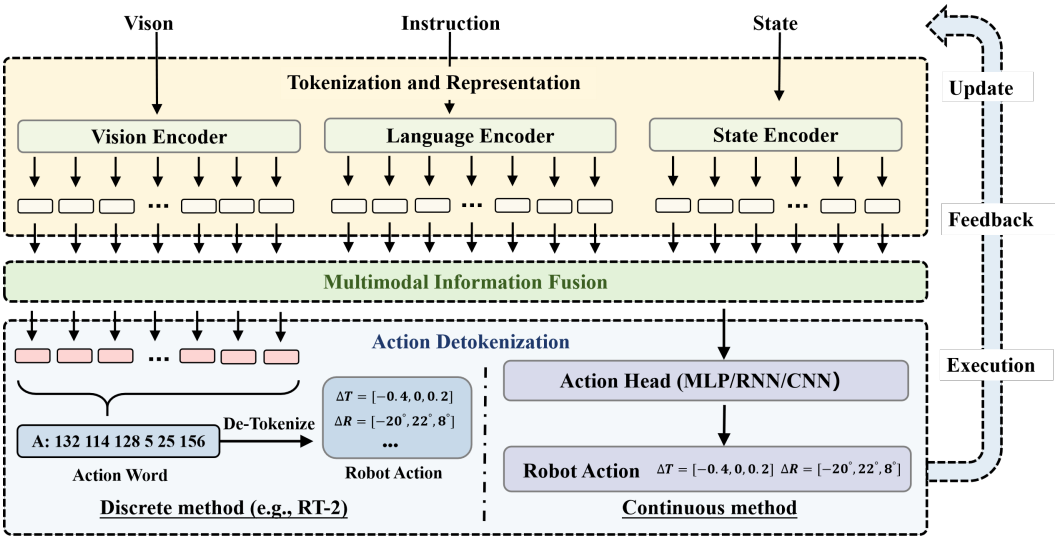


Fig. 10. 视觉-语言-动作模型。

(1) 分词和表示。VLA 模型使用四种标记类型：视觉、语言、状态和动作，以对多模态输入进行编码，从而生成情境感知的动作。视觉标记和语言标记将环境场景和指令编码为嵌入，形成任务和情境的基础。状态标记捕捉代理的物理配置，包括关节位置、力矩、夹持器状态、末端执行器姿态和物体位置。动作标记是基于先前标记自回归生成的，代表低级控制信号（如关节角度、力矩、轮子速度）或高级运动原语（如“移动到抓取姿势”，“旋转手腕”），使 VLA 模型能够作为语言驱动的策略生成器。(2) 多模态信息融合。视觉标记、语言标记和状态标记通过跨模态注意机制融合为用于决策的统一嵌入，该机制通常在变压器架构内实现。此机制动态地权衡各模态的贡献，使 VLA 模型能够在任务情境下共同推理对象语义、空间布局和物理约束。(3) 动作去标记化。融合后的嵌入向量接下来被传递给一个自回归解码器，该解码器通常在变压器架构中实现，以生成一系列动作标记，这些标记对应于低级控制信号或高级运动原语。动作生成可以是离散的或连续的。在离散动作生成中，模型从预定义的动作集中进行选择，例如特定的运动原语或离散化的控制信号，这些被映射为可执行命令。在连续动作生成中，模型输出细粒度的控制信号，通常通过最终的 MLP 层从连续分布中抽样，以实现精确的操控或导航。这些动作标记通过映射到可执行的控制命令进行去标记化，然后传递给执行循环。该循环反馈更新的状态信息，使得 VLA 模型能够动态适应扰动、物体移动或遮挡的实时变化。

机器人变压器 2（RT-2）[222] 是一个著名的 VLA 模型。它利用视觉变压器（ViT）[39] 进行视觉处理，并利用 PaLM 整合视觉、语言和机器人状态信息。特别是，RT-2 将动作空间离散化为八个维度（包括 6 自由度末端执行器位移、夹爪状态和终止命令）。每个维度除终止命令外分为 256 个离散间隔，并嵌入到 VLM 词汇表中作为动作标记。训练期间，RT-2 采用两阶段策略：首先通过互联网规模的视觉语言数据进行预训练，以增强语义泛化；然后进行微调以将输入（即机器人摄像头图像和任务文本描述）映射到输出（即动作词标记序列，

如“1 128 91 241 5 101 127 255”）。训练后的 VLA 模型可以根据视觉语言输入以自回归方式生成动作词，并通过预定义的映射表解码为具体的动作序列。通过将动作建模为“语言”，RT-2 利用大型模型的能力以丰富的语义知识增强低级动作命令。

尽管 VLA 端到端决策架构功能强大，但其存在显著局限性，这限制了其在复杂体现任务中的表现。首先，实时闭环机制导致 VLA 模型对视觉和语言输入的扰动高度敏感，其中视觉噪声（例如，遮挡或背景杂乱）可能使动作输出不稳定，从而影响任务的可靠性。此外，对二维感知的依赖限制了模型解释复杂三维空间关系的能力。其次，动作生成过程通常依赖于输出层的简单策略网络，这难以满足高精度和动态变化任务的要求，导致次优轨迹。第三，训练 VLA 模型需要高计算资源，导致高部署成本和可伸缩性挑战。为了应对这些问题并推动 VLA 在复杂场景中的适用性，研究人员提出了一些增强措施。我们将其分类为：感知能力增强（解决第一个问题）、轨迹动作优化（解决第二个问题）和训练成本降低（解决第三个问题），如图 ?? 所示。

为了提高感知能力，BYO-VLA [66] 通过实现运行时观察干预机制来优化标记化和表示组件，该机制利用自动图像预处理来过滤来自遮挡物体和杂乱背景的视觉噪声。TraceVLA [218] 专注于多模态信息融合组件，引入视觉轨迹提示到跨模态注意力机制。通过将与轨迹相关的数据与视觉、语言和状态标记结合，TraceVLA 增强了时空感知能力，从而能够准确预测动作轨迹。BYO-VLA 提高了输入质量，而 TraceVLA 则在融合过程中完善了动态信息的整合。对于三维感知，3D-VLA [215] 结合了三维大模型和基于扩散的世界模型来处理点云和语言指令。它生成语义场景表示并预测未来的点云序列，提高了三维对象关系理解能力，从而在复杂的三维环境中超越了二维 VLA 模型。SpatialVLA [136] 进一步强调了机器人分类任务中对空间理解的问题。它提出了 Ego3D 位置编码，以将三维信息直接注入输入观测中，并采用自适应行动方案以提高机器人在不同环境中的适应能力。

3.1.1 **轨迹动作优化**. 离散动作空间限制了未定义或高精度动作的表达。通过扩散模型对复杂的机器人行为进行建模，扩散增强方法可以提供更平滑和更可控的动作。Octo [169] 结合了 Transformer 和扩散模型来生成机器人动作。它通过 Transformer 处理多模态输入，提取视觉-语言特征，并使用条件扩散解码器根据这些特征迭代优化动作序列，从而生成平滑且精确的轨迹。通过模块化设计和高效微调，Octo 仅用少量的任务特定数据就实现了跨任务泛化。Diffusion-VLA [185] 将语言模型与扩散策略解码器结合为一个统一框架。它使用自回归语言模型解析语言指令并生成初步任务表示，然后将其输入扩散策略解码器，通过逐步去噪过程优化动作序列。Diffusion-VLA 在整个框架中进行端到端训练，共同优化语言理解和动作生成。扩散过程修正了每一步自回归输出中的不连续性，确保动作轨迹的平滑性和鲁棒性。与 Octo 相比，Diffusion-VLA 具有更高的计算成本，但更适合需要深度语义-动作融合的复杂任务。

3.1.2 **训练成本降低**. VLA 模型在复杂任务中需要高计算成本，这在资源受限的具身平台上受到限制。为了降低训练成本，研究人员提出了优化方法，以提高推理速度、数据效率和实时性能，同时保持任务性能。π_0 [16] 利用流匹配来表示复杂的连续动作分布。与扩散模型中使用的多步采样相比，流匹配通过连续流场建模优化动作生成过程，从而减少计算开销并提高实时性能。相比 Diffusion-VLA [185] 和 Octo [169]，计算效率和控制精度的改进使 π_0 更适合资源受限的具身应用，尤其是需要高精度连续控制的任务。此外，TinyVLA [222] 通过设计轻量级多模态模型和扩散策略解码器，实现了推理速度和数据效率的显著提升。OpenVLA-OFT [83] 使用并行解码代替传统的自回归生成，在单次前向传播中生成完整的动作序列，而不是逐个生成，从而显著减少了推理时间。

3.2 主流 VLA 模型

最近出现了大量的 VLA 模型，具有各种架构和功能。为了更好地理解和部署，我们在表 2 中总结并比较了它们的架构、贡献和能力增强。

Table 2. 主流的 VLA 模型（P：感知，A：轨迹动作，C：训练成本）。

Model	Architecture	Contributions	Enhancements		
			P	A	C
RT-2 [222] (2023)	• 视觉编码器：ViT22B/ViT-4B • 语言编码器: PaLIX/PaLM-E • 动作解码器：符号调谐	Pioneering large-scale VLA, jointly fine-tuned on web-based VQA and robotic datasets, unlocking advanced emergent functionalities.	×	√	√
Seer [55] (2023)	• 视觉编码器：视觉骨干网络 • 语言编码器：基于 Transformer • 动作解码器：自回归动作预测头	Efficiently predict future video frames from language instructions by extending a pretrained text-to-image diffusion model.	√	×	√
Octo [169] (2024)	• 视觉编码器：CNN • 语言编码器: T5-base • 动作解码器：扩散变压器	First generalist policy trained on a massive multi-robot dataset (800k+ trajectories). A powerful open-source foundation model.	×	√	×
Open-VLA [85] (2024)	• 视觉编码器：DINOv2 + SigLIP • 语言编码器：Prismatic-7B • 动作解码器：符号调谐	An open-source alternative to RT-2, superior parameter efficiency and strong generalization with efficient LoRA fine-tuning.	×	×	√
Mobility-VLA [34] (2024)	• 视觉编码器：长上下文 ViT + 目标图像编码器 • 语言编码器：基于 T5 的指令编码器 • 动作解码器：混合扩散 + 自回归集合	Leverages demonstration tour videos as an environmental prior, using a long-context VLM and topological graphs for navigating based on complex multimodal instructions.	√	√	×
Tiny-VLA [187] (2025)	• 视觉编码器：具有低延迟编码的 FastViT • 语言编码器：紧凑型语言编码器（128 维） • 动作解码器：扩散策略解码器（5000 万参数）	Outpaces OpenVLA in speed and precision; eliminates pretraining needs; achieves 5x faster inference for real-time applications.	×	×	√
Diffusion-VLA [185] (2024)	• 基于 Transformer 的视觉编码器用于上下文感知 • 语言编码器：带有下一个标记预测的自回归推理模块 • 用于鲁棒动作序列生成的扩散策略头	Leverage diffusion-based action modeling for precise control; superior contextual awareness and reliable sequence planning.	×	√	×
Point-VLA [95] (2025)	• 视觉编码器：CLIP + 三维点云 • 语言编码器：Llama-2 • 动作解码器：具有空间标记融合的 Transformer	Excel at long-horizon and spatial reasoning tasks; avoid retraining by preserving pretrained 2D knowledge	√	×	×
VLA-Cache [197] (2025)	• 视觉编码器：带有记忆缓冲区的 SigLIP • 语言编码器：Prismatic-7B • 动作解码器：具有动态令牌重用的 Transformer	Faster inference with near-zero loss; dynamically reuse static features for real-time robotics	×	×	√

Table 2. 主流的 VLA 模型（P：感知，A：轨迹行动，C：训练成本）（续）。

Model	Architecture	Contributions	Enhancements		
			P	A	C
π_0 [16] (2024)	• 视觉编码器：PaliGemma VLM 主干 • 语言编码器：PaliGemma（多模态） • 动作解码器：流匹配	Employ flow matching to produce smooth, high-frequency (50Hz) action trajectories for real-time control.	×	√	√
π_0 Fast [132] (2025)	• 视觉编码器：PaliGemma VLM 骨架 • 语言编码器：PaliGemma（多模态） • 动作解码器：带有 FAST 的自回归 Transformer	Introduces an efficient action tokenization scheme based on the Discrete Cosine Transform (DCT), enabling autoregressive models to handle high-frequency tasks and significantly speeding up training.	×	√	√
Edge-VLA [23] (2025)	• 视觉编码器：SigLIP + DINOv2 • 语言编码器：Qwen2（0.5B 参数） • 动作解码器：联合控制预测（非自回归）	Streamlined VLA tailored for edge devices, delivering 30–50Hz inference speed with OpenVLA-comparable performance, optimized for low-power, real-time deployment.	×	×	√
OpenVLA-OFT [83] (2025)	• 视觉编码器：SigLIP + DINOv2（多视图） • 语言编码器：Llama-2 7B • 动作解码器：使用动作分块和 L1 回归的并行解码	An optimized fine-tuning recipe for VLAs that integrates parallel decoding and a continuous action representation to improve inference speed and task success.	×	√	√
Spatial-VLA [136] (2025)	• 视觉编码器：来自 PaLiGemma2 4B 的 SigLIP • 语言编码器：PaLiGemma2 • 动作解码器：自适应动作网格和自回归变换器	Enhance spatial intelligence by injecting 3D information via 'Ego3D Position Encoding' and representing actions with 'Adaptive Action Grids'.	√	√	×
MoLe-VLA [208] (2025)	• 视觉编码器: 多阶段 ViT 与 STAR 路由器 • 语言编码器：CogKD 增强型 Transformer • 动作解码器：具有动态路由的稀疏 Transformer	A brain-inspired architecture that uses dynamic layer-skipping (Mixture-of-Layers) and knowledge distillation to improve efficiency.	×	×	√
DexGrasp-VLA [219] (2025)	• 视觉编码器：以对象为中心的空间 ViT • 语言编码器：具有抓取序列推理的 Transformer • 动作解码器：用于抓取姿势生成的扩散控制器	A hierarchical framework for general dexterous grasping, using a VLM for high-level planning and a diffusion policy for low-level control.	×	√	×
Dex-VLA [186] (2025)		A large plug-in diffusion-based action expert and an embodiment curriculum learning strategy for efficient cross-robot training and adaptation.	×	√	×

3.3 分层决策与端到端决策

层次化和端到端代表了实现具身智能自主决策的两种不同范式，每一种都有其独特的设计理念、实现策略、性能特征和应用领域。我们在此对它们进行比较，如表 3 所示，该表概述了在架构、性能、可解释性和泛化能力等方面的关键差异。

层次结构将决策过程分解为多个模块，每个模块处理感知、计划、执行和反馈的特定方面。其核心思想是将复杂任务分解为可管理的子任务，以增强可调试性、优化和维护。层次结构在整合领域知识（例如，物理约束、规则）方面表现卓越，为具体化任务提供高度的可解释性和可靠性。但它们的局限性也很明显。模块的分离可能会因不当协调而导致次优解决方案，特别是在动态复杂环境中。人工任务分解可能阻碍对未见场景和任务的适应性。

端到端架构采用大规模神经网络，例如 VLA，将多模态输入直接映射到动作上，而无需模块化分解。VLA 通常建立在大型多模态模型之上，并在海量数据集上进行训练，能够同时实现视觉感知、语言理解和动作生成。由于高度集成的架构，VLA 减少了跨模块的错误积累，并通过端到端优化实现高效学习。随着在大规模多模态数据集上的训练，VLA 在结构化环境中的复杂任务中具有很强的泛化能力。然而，VLA 的黑箱性质降低了解释性，使得分析决策过程变得困难。VLA 的性能高度依赖于训练数据的质量和多样性。端到端训练的计算成本也很高。

Table 3. **分层和端到端决策范式的比较。**

Aspect	Hierarchical	End-to-End
Architecture	• 感知：专用模块（例如，SLAM，CLIP） • 高级规划: 结构化, 语言, 程序 • 低级执行：预定义的技能列表 • 反馈：LLM 自我反思、人类、环境	• 感知：整合在分词中 • 规划：通过 VLA 预训练隐式进行 • 动作生成：基于扩散的解码器进行自回归生成 • 反馈：闭环周期中的固有属性
Performance	• 在结构化任务中可靠 • 在动态环境中受限	• 在复杂、开放性任务中具有强大的概括能力 • 依赖于训练数据
Interpretability	• 高，具有清晰的模块化设计	• 由于神经网络的黑箱性质，处于低水平
Generalization	• 局限于依赖人为设计的结构	• 强，由大规模预训练驱动 • 对数据缺口敏感
Real-time	• 低模块间通信可能会在复杂场景中引入延迟	• 高效、直接的感知到行动映射可最小化处理开销
Computational Cost	• 适中，有独立模块优化，但有协调开销	• 高，需要大量资源进行训练
Application	• 适用于工业自动化、无人机导航、自动驾驶	• 适用于家用机器人、虚拟助手、人机协作
Advantages	• 高可解释性 • 高可靠性 • 容易整合领域知识	• 无缝多模态集成 • 在复杂任务中有效 • 最小误差累积
Limitations	• 由于模块协调问题而次优 • 对非结构化环境的适应能力较低	• 低可解释性 • 对训练数据的高度依赖 • 高计算成本 • 在分布外情境中的低泛化能力

4 具身学习

体现学习旨在使智能体能够在与环境互动的过程中获得复杂技能并提升其能力 [216] 。通过持续学习和优化技能，智能体可以实现精确决策和实时适应。这种能力可以通过多种学习策略的协调来实现，如图 11 所示。模仿学习允许智能体快速获取初始策略，迁移学习 [141] 促进知识在不同任务间的传递，元学习 [45] 使智能体学会如何学习，强化学习 [9] 则通过与环境的持续互动来优化策略。然而，这些学习方法在体现智能中仍面临显著的技术挑战。模仿学习难以捕捉复杂行为，而强化学习常常受到设计有效奖励函数复杂性的阻碍。近年来，Transformer 和大模型的出现激励研究人员探索将大模型与学习方法整合以克服这些限制。在本节中，我们首先描述体现学习的过程和常用学习方法，然后详细阐述模仿学习和强化学习，并探讨大模型如何在体现智能中增强这些方法。

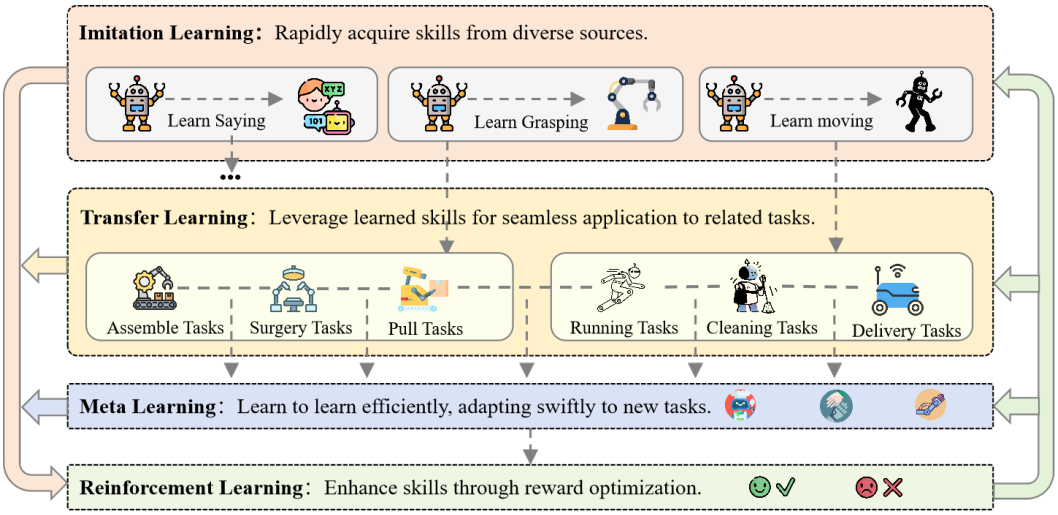


Fig. 11. 具身学习：过程和方法论。

4.1 具身学习方法

具身代理应该能够在其整个生命周期中获取新知识和学习新任务，而不是依赖初始训练数据集 [216] 。这种能力对于现实世界的复杂性和多样性至关重要，在现实世界中，新任务和挑战频繁出现。具身学习可以建模为一个目标条件部分可观测马尔可夫决策过程，定义为一个 8 元组 $(S, A, G, T, R, \Omega, O, \gamma)$ ，其中

- S 是环境状态的集合。每个状态编码多模态信息，如文本描述、图像或结构化数据。
- A 是一组动作。每个动作代表一个指令或命令，通常用自然语言表示。
- G 是可能目标的集合。$g \in G$ 指定了一个特定的目标，例如，购买一台笔记本电脑。
- $T(s'|s, a)$ 是状态转移概率函数。对于每个状态-动作对 (s, a) ，$T(\cdot)$ 定义了下一个状态 $s' \in S$ 的概率分布。
- $R : S \times A \times G \to R$ 是目标条件奖励函数，它用于评估在状态 s 中，动作 a 在多大程度上推动了目标的实现。对于每个三元组 (s, a, g) ，奖励可以是数值型（例如，一个分数）或文本型（例如，干得好），为目标提供了互动反馈。
- Ω 是一组观测值，可能包括文本、视觉或多模态数据，表示代理对于状态的部分观察。
- $O(o'|s', a)$ 是观测概率函数，定义了在通过动作 a 转换到状态 s' 后观测到 $o' \in \Omega$ 的概率。

- $\gamma \in [0,1)$ 是折扣因子，用于平衡即时和长期奖励。它仅在奖励为数值时适用。

该表述捕捉了现实场景中的复杂性，其中智能体在部分可观察的随机动态下运行。在时间 t ，智能体接收到观测 $o_t \in \Omega$ 和目标 $g \in G$ 。智能体根据策略 $\pi(a_t|o_t, g)$ 选择行动 $a_t \in A$ 。执行行动后，环境状态转变为 $s_{t+1} \sim T(s'|s_t, a_t)$ ，产生观测 $o_{t+1} \sim O(o'|s_{t+1}, a_t)$ 和奖励 $R(s_{t+1}, a_t, g)$ 。对于端到端的决策，VLA 模型直接编码策略 $\pi(a|o, g)$ ，处理多模态观测 $o \in \Omega$ 并生成行动 $a \in A$ 。对于层次化的决策，高级智能体通过增强的 LLM 策略 $\pi_{high}(g_{sub}|o, g)$ 生成具有上下文感知的子目标 g_{sub} ，然后低级策略 $\pi_{low}(a|o, g_{sub})$ 将子目标映射为行动 $a \in A$ 。低级策略 $\pi_{low}(a|o, g_{sub})$ 可以通过模仿学习或强化学习来学习。学习到的策略嵌入在模型的层次结构中，并在训练期间进行微调以处理特定任务，例如导航、操控、人机交互。

对于具身智能，模仿学习、强化学习、迁移学习和元学习在使代理能够在复杂现实环境中行动中都扮演着重要角色。每种学习方法都解决了不同的挑战。模仿学习 [193] 通过模仿专家或视频示范让代理学习有效策略，对于机器人操作等具有高质量数据可得性的任务非常高效。但它对多样化示范的依赖限制了在新场景中的适应能力。强化学习 [129] 在动态环境中通过试错交互，由奖励函数引导而表现优异。然而，设计合适的奖励函数是一项挑战，并且强化学习需要高计算资源。迁移学习 [141] 通过在相关任务间转移知识来增强学习效果，非常适合技能的重用。然而，当任务显著不同时，它存在负迁移的风险 [170] 。元学习 [58] 专注于学习如何学习，使得在很少数据的情况下能够快速适应新任务。但它需要在多样化任务上的广泛预训练。表 4 简要总结和比较了这些方法在具身 AI 中的应用。

Table 4. **关于具身人工智能的学习方法比较**。

Methods	Strengths	Limitations	Applications
Imitation Learning	• 通过模仿专家示范进行快速政策学习 • 适用于高质量数据的任务	• 依赖于多样化且高质量的示范 • 对新任务或稀疏数据场景的适应能力有限	• 机器人操作 • 结构化导航 • 在专家指导下的人机交互
Reinforcement Learning	• 通过试错法在动态不确定环境中优化策略 • 在具有明确奖励信号的任务中表现出色	• 需要大量样本和计算资源 • 对奖励函数和折扣因子敏感	• 自主导航 • 自适应人机交互 • 动态任务优化
Transfer Learning	• 通过在相关任务之间转移知识来加速学习 • 增强相关任务中的泛化能力	• 当任务差异显著时，存在负迁移风险 • 需要任务相似性以实现有效学习	• 穿越多样化环境的导航 • 使用共享结构进行操作 • 跨任务技能重用
Meta-Learning	• 使用极少数据快速适应新任务 • 理想的多样化具体任务	• 需要广泛的预训练和大型数据集 • 建立一个通用的元政策是资源密集型的	• 在不同任务和环境中快速适应导航、操作或交互

4.1.1 **模仿学习** . 模仿学习是体现学习中的一种关键方法。它通过模仿专家演示使智能体能够学习策略，从而快速获取针对目标导向任务的决策策略 [193] 。训练是通过专家状态-动作对的数据集进行监督的 (s, a) 。其目标是通过最小化专家动作的负对数似然来学习一种策略 $\pi(a|s)$ ，以密切模仿专家的行为。因此，其目标函数可以定义为：

$$\mathcal{L}(\pi) = -\mathrm{E}_{\tau \sim \mathrm{PD}}[\log \pi(\mathrm{a}|s)] \tag{1}$$

其中 D 是专家演示的集合。每个演示 τ_i 由一系列长度为 L 的状态-动作对 (s_t, a_t) 组成：

$$\tau_i = [(s_1, a_1), \cdots, (s_t, a_t), \cdots, (s_L, a_L)] \tag{2}$$

在连续动作空间中，策略 $\pi(\cdot)$ 通常被建模为高斯分布，目标函数通过预测和专家动作之间的均方误差（MSE）来近似。模仿学习的样本效率很高，因为它避免了大量的试错，但它高度依赖于演示数据的质量和覆盖范围，在未见过的场景中存在困难。结合模仿学习与强化学习的混合方法可以通过使用模仿学习初始化策略并使用强化学习加以改进来解决这一限制，从而增强对未见情形的鲁棒性。

4.1.2 **强化学习**. 强化学习目前是具身学习中的主流方法。它使智能体能够通过在环境中通过试错进行交互来学习策略，使得其非常适合动态和不确定的环境 [129] 。在每个时间步 t ，智能体观察一个状态 s 并根据其策略 $\pi(a|s)$ 选择一个动作 a 。在动作执行后，智能体从奖励函数 $R(s,a,g)$ 获得一个奖励 r ，环境根据状态转移概率 $T(s'|s,a)$ 转变为新的状态 s' ，生成一个观测 $o' \sim O(o'|s',a)$ 。强化学习的目标函数是最大化期望累积奖励：

$$\mathcal{J}(\pi) = E_{\pi,T,O}\left(\sum_{t=0}^{\infty} \gamma^t R(s,a,g)\right) \tag{3}$$

其中 $\gamma \in [0,1)$ 是平衡即时和长期奖励的折扣因子。强化学习擅长优化复杂任务的策略，但需要广泛的探索，这在计算上是代价高昂的。结合模仿学习和强化学习的混合方法可以改善这一问题，其中模仿学习提供初始策略以减少探索，而强化学习通过与环境的交互来精细化这些策略。

在从头训练需要大量样本和时间的情况下，可以应用迁移学习来减轻工作量。它允许智能体使用源任务的知识来加速相关目标任务的学习。通过从源任务转移学习到的策略、特征或表示，智能体提高了在目标任务上的效率和泛化能力。给定一个由状态 $s \in S$ 、动作 $a \in A$ 和策略 $\pi(a|s)$ 定义的源任务，迁移学习将源策略 π_s 适应于具有不同动态或目标的目标任务。其目标是通过使用少量目标任务数据微调策略，来最小化源策略 π_s 和目标策略 π_t 之间的差异。该过程由目标任务的特定任务损失引导，并受到策略对齐的 Kullback-Leibler（KL）散度的约束：其中，θ_t^* 代表目标任务的最优策略参数，θ_s 和 θ_t 分别是源策略和目标策略的参数，D_{KL} 衡量源策略 π_s 和目标策略 π_t 之间的散度，$\mathcal{L}_t$ 是目标任务的特定任务损失，λ 是平衡策略对齐和任务表现的正则化参数。这个过程确保转移的知识与目标任务的状态转移概率 $T(s'|s,a)$ 和奖励函数 $R(s,a,g)$ 对齐。在具身设置中，迁移学习使智能体能够在不同环境和目标中重复使用学会的行为，从而减少训练时间。然而，源任务和目标任务之间的显著差异可能导致负迁移，即由于知识不匹配而导致的性能下降。

元学习也可以用于具身人工智能，使代理能够学习如何学习，从而能够通过少量样本迅速推断出新任务的最优策略。在每个时间步中，代理接收到观测值和目标，并根据适应由状态转移概率和奖励函数定义的任务特定动态的元策略选择动作。其目标是通过最小化任务特定数据的损失来优化跨任务的预期性能。在模型无关元学习（MAML）的背景下，这是通过学习一组初始模型参数来实现的，该参数可以通过最小更新快速适应新任务。具体来说，对于一组任务，MAML 优化元目标如下：其中表示最优元策略参数，是任务特定的损失，是由参数化的模型，是使用学习率进行梯度更新后的任务特定参数，外在优化在适应后最小化任务间的预期损失。元学习可以通过微调预训练模型并使用少量演示或交互使代理能够快速适应新任务。元策略可以嵌入到大型模型中，并在训练过程中进行细化以处理多样化的任务。尽管具有优势，元学习需要大量的预训练和跨任务的大量样本，给建立通用学习策略带来了挑战，特别是当任务在状态空间或动态上差异显著时。

4.2 由大模型赋能的模仿学习

模仿学习的主要目标是使智能体通过模仿示范者的动作来达到专家级别的表现。模仿学习可以通过不同的方法来实现，包括行为克隆 [46] 、逆向强化学习 [125] 、生成对抗模仿学习 [71] 和层级模仿学习 [10] ，每种方法都对策略网络的构建有所贡献。在这些方法中，行为克隆是最重要的一种，它将模仿学习形式化为一个监督回归任务。给定观察 $o \in \Omega$ 和目标 $g \in G$ ，策略网络 π 预测预期的动作 $a \in A$ 。

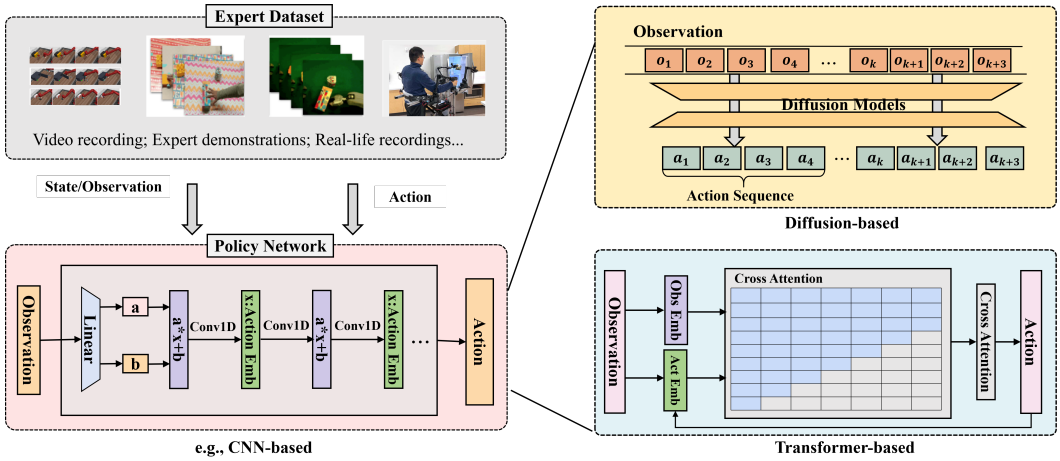


Fig. 12. **模仿学习由扩散模型或 Transformers 驱动**。

策略网络 π 需要准确地将观测 o 和目标 g 映射到动作 a，以确保高拟合度，即使在复杂的、动态的、部分可观察的环境中也是如此。模仿学习不仅仅是复制，它还旨在赋予代理在未见过的状态、目标或环境中泛化的能力。这种泛化能力对于真实世界应用至关重要，例如机器人操作、自动导航和人机交互，在这些情境中，环境动态和任务需求往往与训练情境不同。此外，模仿学习还旨在确保对分布偏移的鲁棒性，即在随机或动态设置中，小的动作预测错误不会积累成对专家轨迹产生显著偏离。最后，模仿学习力求样本效率，试图让代理从有限数量的专家示范中学习有效的策略，从而减少对大量高质量数据集的依赖。

行为克隆在构建鲁棒的策略 [193] 方面仍然有困难。它对高质量专家演示的依赖阻碍了对未见状态或目标的泛化。专家演示通常表现出随机性、多模态性和复杂性，这对于策略网络来说难以捕捉，导致模仿保真度受损和性能下降。最近在大型模型方面的进展显著增强了行为克隆，解决了其固有的局限性。如图 12 所示，大型模型在以下几个方面赋能模仿学习：(1) 使用扩散模型构建策略网络；(2) 使用 Transformer 构建策略网络。

4.2.1 **基于扩散的策略网络**. 扩散模型在处理复杂多模态分布方面非常出色 [35]。它们可以用于生成多样的动作轨迹，从而增强策略的鲁棒性和表现力。最近的研究开始将扩散模型与策略网络结合，以克服传统模仿学习的局限性。Pearce [131] 提出了一种基于扩散模型的模仿学习框架，该框架将扩散模型合并到策略网络中。通过噪声添加和移除的迭代优化专家演示，该框架能够捕捉动作分布的多样性并生成多样的动作序列。DABC [31] 采用两阶段流程，通过扩散模型的赋能来训练策略网络。首先通过行为克隆预训练基础策略网络，然后通过扩散模型细化动作分布的建模。扩散政策 [33] 提出了一种将扩散模型作为视觉驱动机器人物任务的决策模型的策略网络。它使用视觉输入和机器人的当前状态作为条件，采用 U-Net 作为去噪网络，根据视觉特征和状态向量预测去噪步骤，从而生成连续的动作序列。为了增强策略网络的空间感知能力，3D-Diffusion [206] 提出了一种基于 3D 输入的扩散策略框架。它使用简单的 3D 表示作为输入，利用扩散模型生成动作序列，从而通过捕捉空间信息来提高视觉运动策略的泛化能力。与 2D 策略网络相比，3D-Diffusion 能够更好地理解 3D 环境中的几何关系和空间约束。

4.2.2 **基于 *Transformer* 的策略网络**. Transformer 架构能够通过将专家轨迹视为序列数据并利用自注意力机制来建模动作、状态和目标之间的依赖关系，从而增强模仿学习。该端到端方法在中间步骤中减少误差累积，提升策略的一致性和准确性。谷歌的 RT-1 [18] 首次展示了 Transformer 在机器人控制中的潜力。通过结合大规模多样化的数据集（13 万多个轨迹，

700 多个任务）和预训练的视觉语言模型，它显著提高了对未见任务和场景的任务泛化能力。后续工作 RT-Trajectory [54] 引入了“轨迹草图”方法，结合低级视觉线索以增强端到端 Transformer 的任务泛化能力。斯坦福大学的 ALOHA [213] 利用 Transformer 的编码-解码结构，从多视图图像生成机械臂的动作序列，利用低成本硬件实现精确的双臂操作。其后续研究使用动作分块策略预测多步动作序列，显著提高了长期任务的稳定性和一致性。Mobile ALOHA [50] 将原有任务扩展到全身协调的移动操作任务，引入了移动平台和远程操作界面以处理更复杂的双臂任务。在 3D 空间操作中，HiveFormer [213] 和 RVT [52] 利用多视图数据和 CLIP 进行视-语言特征融合，直接预测 6D 抓取姿势，在 RLBench 和现实世界机器人臂任务上取得了最先进的性能，突显了 Transformer 在复杂空间建模中的优势。为了抓取可变形物体（例如，织物或软材料），Man 提出了结合视觉和触觉反馈的 Transformer 框架，通过探索性动作优化抓取参数。谷歌的 RoboCat [17] 采用跨任务、跨实体的具体模仿学习，结合 VQ-GAN [44] 以标记化视觉输入，利用决策 Transformer 预测动作和观察结果，仅需少量样本即实现快速的策略泛化。RoboAgent [15] 采用类似的编码-解码结构，融合视觉、任务描述和机器人状态以最小化动作序列预测误差。CrossFormer [38] 提出了一种基于 Transformer 的模仿学习架构用于跨实体任务，该架构在大规模专家数据上进行训练，以统一处理操作、导航、移动和空中任务，展示了多任务学习的潜力。

4.3 由大型模型增强的强化学习

通过与环境的交互，强化学习 [9] 使得智能体能够开发出最优的控制策略，适应多样的未知情境，在动态环境中保持稳健性，并从有限的数据中学习，从而在现实世界中实现复杂的任务。最初，强化学习基于基本技术，如策略搜索和价值函数优化，典型例子包括 Q-learning [183] 和状态动作回报状态动作（SARSA）[153] 。随着深度学习的盛行，强化学习与深度神经网络相结合，被称为深度强化学习（DRL）。DRL 使得智能体能够从高维输入中学习复杂的策略，取得了显著的成就，例如 AlphaGo [163] 和深度 Q 网络（DQN）[120] 。DRL 使得智能体无需明确的人为干预即可在新环境中自主学习，从而允许从游戏到机器人控制及其他领域的广泛应用。随后的进展进一步改善了学习效果。近端策略优化（PPO）[155] 通过截取的概率比率提高了策略优化的稳定性和效率。软演员-评论家（SAC）[60] 通过结合最大熵框架改善了探索性和稳健性。

尽管取得了这些进展，强化学习在构建策略网络 π 和设计奖励函数 $R(s,a,g)$ 方面仍然面临限制。大型模型的最新进展使强化学习在以下几个方面得到了增强：（1）改进奖励函数设计；（2）通过对复杂动作分布进行建模来优化策略网络的构建。它们如图 13 所示。

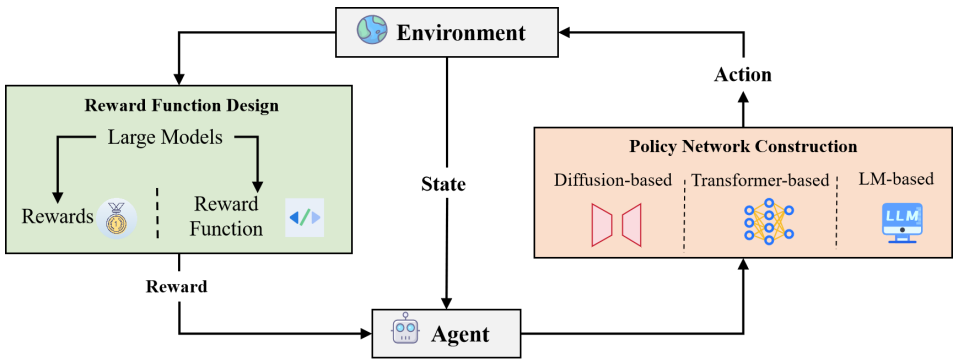


Fig. 13. 由大型模型增强的强化学习。

4.3.1 **奖励函数设计**. 设计奖励函数一直是强化学习的一个挑战 [43] ，因为它的复杂性和任务特定的特性。传统的奖励函数是由领域专家手动设计的，需要全面考虑诸如任务完成、

能耗、安全性以及每个因素的权重等因素，这相当困难。手动设计通常导致稀疏或缩放不佳的奖励，引发诸如奖励欺骗等问题，即智能体利用非预期信号来最大化奖励，却未能实现预期目标。

大型模型提供了一种有前景的解决方案，通过生成（1）奖励信号 r 或（2）奖励函数 $R(s,a,g)$，减少对人工设计的依赖，同时捕捉复杂的多模态反馈。Kwon 等人和 Language to Rewards (L2R) [204] 分别引入了零样本和少量样本的方法，利用 GPT-3 直接从文本行为提示中生成奖励信号，将高层次目标转化为硬件特定的控制策略。然而，它们的稀疏奖励限制了在复杂任务中的使用，成功的生成严重依赖于精确的提示或特定的模版。Text2Reward [194] 通过从环境描述和示例生成密集的可解释的 Python 奖励函数来改进这一点，并通过人类反馈迭代地优化这些函数，在机器人操作和运动任务中实现了高成功率。Eureka [110] 利用 GPT-4 从任务和环境提示中创建密集奖励。它通过采用一种自动的迭代策略来优化奖励函数，减轻了 Text2Reward 对人类反馈的依赖，从而超越了人类设计的奖励。此外，Auto MC-Reward [96] 在 Minecraft 中实现了全自动化，采用多阶段流水线，其中奖励设计者生成奖励信号，验证者确保质量，轨迹分析器通过失败驱动的迭代优化奖励。Auto MC-Reward 显著提升了效率，但其领域特定的重点限制了其相对于 Eureka 和 Text2Reward 的泛化能力。

4.3.2 **策略网络构建**. 离线强化学习从预先收集的数据集中学习 [91] 最优策略，而无需在线交互。但对静态数据集的依赖可能导致数据集中不存在的动作产生误差。策略正则化可以通过约束行为策略的偏差来减轻这个问题。但策略表达能力的局限性和次优的正则化方法可能导致性能不足。为了增强离线强化学习的表达能力和适应性，研究人员提出利用（1）扩散模型，（2）基于 Transformer 的架构，以及（3）LLMs 来增强策略网络的构建，如图 14 所示。

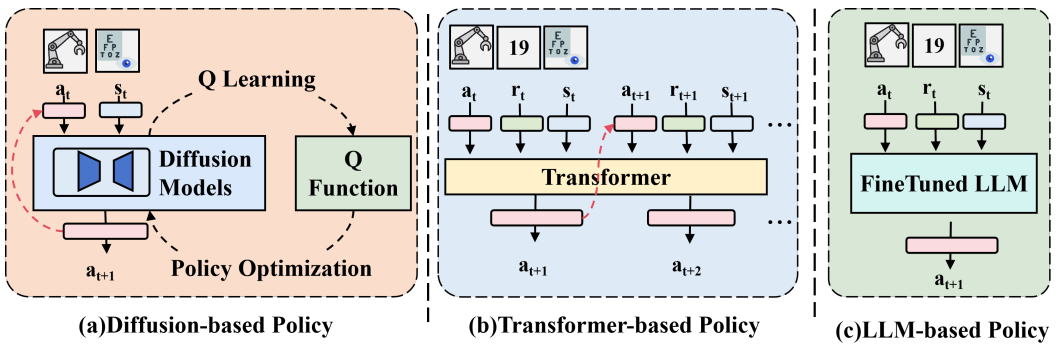


Fig. 14. **由大型模型赋能的策略网络构建**。

使用扩散模型构建策略网络. 扩散模型 [35] 通过迭代噪声添加和去噪来模拟复杂的动作分布，从而增强策略的表达能力。Diffusion-QL [182] 将扩散模型作为基础策略，建模动作分布，并在 Q 学习框架中训练以最大化价值函数目标。该方法生成高奖励策略，适合离线数据集中多峰或非标准动作分布。然而，扩散模型需要大量的去噪步骤才能从完全噪声的状态中生成动作。为减少此工作量，EDP [82] 引入了一种高效的采样方法，该方法在一步中从中间噪声状态中重建动作，显著减少了计算开销。EDP 可以与各种离线强化学习框架集成，在保持策略表达能力的同时提高采样效率。

基于 Transformer 的架构利用自注意力机制来捕捉轨迹中的长期依赖性，从而提高策略的灵活性和准确性。Decision Transformer [29] 将离线强化学习重新构建为一个条件序列建模问题，将状态-动作-奖励轨迹视为序列输入，并应用监督学习从离线数据集中生成最优动作。在此基础上，Prompt-DT [196] 通过结合提示工程，在少样本场景中通过任务特定编码

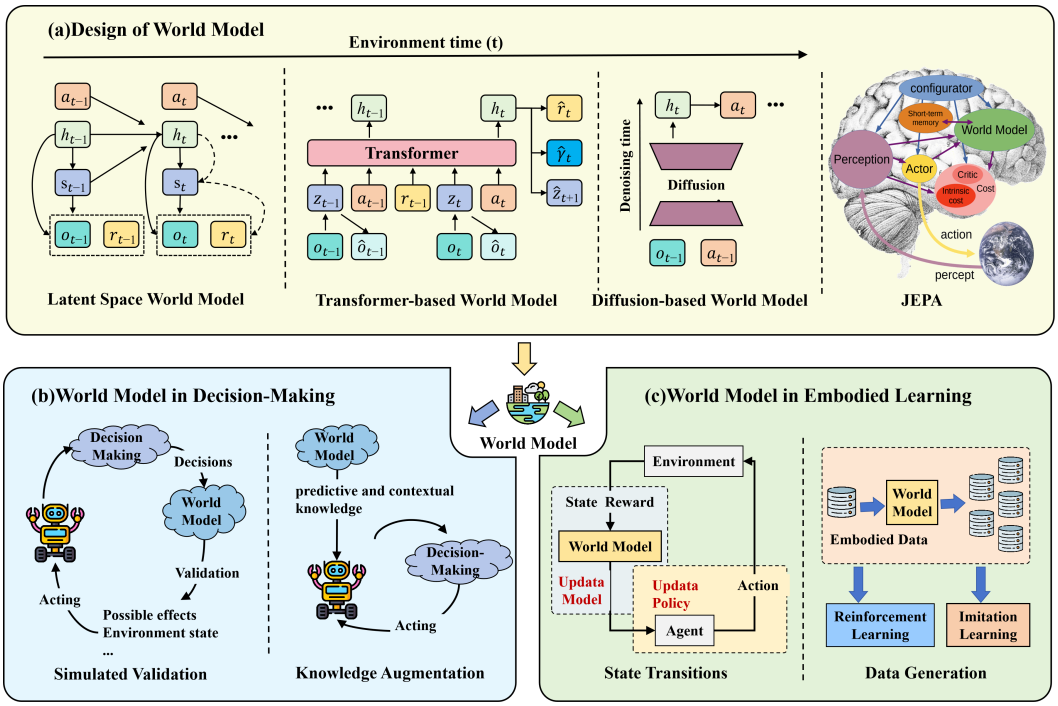


Fig. 15. 世界模型及其在决策和具身学习中的应用。

的轨迹提示来指导新任务的动作生成，从而增强泛化能力。为了提高动态环境中的适应性，Online Decision Transformer (ODT) [217] 在离线强化学习中预训练 Transformer 以学习序列生成，然后通过在线强化学习交互进行微调。Q-Transformer [28] 将 Transformer 的序列建模与 Q 函数估计结合起来，以自回归方式学习 Q 值来生成最优动作。在多任务离线强化学习中，Gato [147] 采用基于 Transformer 的序列建模方法，但它严重依赖于数据集的最优性，并因参数庞大而导致高训练成本。

基于 Transformer 的序列建模能力，LLM 引入了一种新的范式，通过利用预训练知识简化离线强化学习任务。GLAM [26] 使用 LLM 作为策略代理，为语言定义的任务生成可执行的动作序列，这些序列通过 PPO 和上下文记忆在线优化，以提高长时间规划中的序列一致性。LaMo [158] 以 GPT-2 为基础策略，经过 LoRA 微调以保留先验知识，将状态-动作-奖励序列转换为语言提示，以生成与任务一致的策略。Reid [148] 利用预训练的 BERT 探索 LLM 的可迁移性，微调为特定任务并通过外部知识库增强。在 D4RL 基准上的评估 [49] 显示 Reid 在超越 Decision Transformer 的同时减少了训练时间，展示出 LLM 在离线强化学习中的效率。

5 世界模型

世界模型作为环境的内部模拟或表示。有了世界模型，智能系统可以预测未来状态，理解因果关系，并做出合理的决策，而无需仅仅依赖于现实世界的交互，这既昂贵又常常不可行。通过提供丰富的认知框架，世界模型促进了在复杂动态环境中更高效的学习、决策和适应性，从而增强了代理执行复杂任务的能力。在本节中，我们探讨世界模型的设计，并研究它们如何促进决策和具身学习。

5.1 世界模型的设计

世界模型的概念可以追溯到强化学习 [221] 。传统的强化学习依赖于代理-环境的反复交互，导致计算成本高，因此在数据稀缺或复杂场景中不切实际。世界模型使得代理可以在模拟环境中学习，而不是仅仅通过反复交互学习行为。这种方法在数据稀缺或复杂场景中特别有价值。在设计方面，目前的世界模型可以分为四种类型：潜在空间世界模型、基于 Transformer 的世界模型、基于扩散的世界模型和联合嵌入预测架构，如图 15 的上部分所示。

潜在空间世界模型由递归状态空间模型（RSSM）[59, 61] 表示，能够实现潜在空间中的预测。RSSM 从像素观测中学习动态环境模型，并在编码的潜在空间中规划动作。通过将潜在状态分解为随机和确定性部分，RSSM 同时考虑环境的确定性和随机性因素。由于 RSSM 在机器人连续控制任务中的出色表现，出现了许多基于 RSSM 的工作。PlaNet [63] 采用结合门控循环单元（GRU）和卷积变分自编码器（CVAE）的 RSSM，利用 CNN 进行潜在动力学和模型预测控制。Dreamer [62] 通过从潜在表示中学习动作网络和价值网络促进了这一点。Dreamer V2 [64] 进一步使用演员-评论算法完全从世界模型生成的想象序列中学习行为，在 Atari 200M 基准上达到了可与人类玩家相比的性能。Dreamer V3 [65] 通过 symlog 预测、层归一化和通过指数移动平均的归一化回报提高了稳定性，在连续控制任务中表现优于专用算法。

5.1.1 **基于 *Transformer* 的世界模型** . 潜在空间世界模型通常依赖于 CNN 或递归神经网络 (RNN)，因此在高维、连续或多模态环境中操作时面临挑战。基于 Transformer 的世界模型提供了一种强大的替代方法。它们利用注意力机制来建模多模态输入，克服了 CNN 和 RNN 的局限性，在复杂的记忆交互任务中表现出卓越的性能。IRIS [119] 是最早将 Transformer 应用于世界模型之一，其中代理在基于自回归 Transformer 的世界模型中学习技能。IRIS 使用矢量量化变分自编码器（VQ-VAE）对图像进行标记，并采用自回归 Transformer 预测未来的标记，在数据量少的 Atari 100k 设置中表现出色。谷歌的 Genie [22] 基于时空 Transformer [195] ，通过自监督学习在庞大的未标记互联网视频数据集上进行了训练，优于传统的 RSSM。Genie 为可操作的、生成的、交互的环境提供了一种新的范式，突显了 Transformer 的变革潜力。TWM [151] 提出了一种基于 Transformer-XL 的世界模型。它将 Transformer-XL 的片段级递归机制迁移到世界模型中，从而能够捕捉环境状态之间的长期依赖关系。为了进一步提高效率，TWM 在潜在想象中训练一个无模型的代理，避免在运行时进行完整推理。STORM [211] 使用随机 Transformer，因此不依赖于在 Atari 100k 基准上进行前瞻搜索。它将状态和动作融合为单一的标记，提高了训练效率，并在 Atari 100k 基准上匹配了 Dreamer V3 的性能。这些基于 Transformer 的世界模型将状态、动作和观测离散化为序列，利用自注意力来捕捉长期依赖关系，显著提高了预测准确性、样本效率以及在不同任务中的适应性。

5.1.2 **基于扩散的世界模型** . 基于扩散的世界模型，以 OpenAI 的 Sora 为代表，在原始图像空间中生成预测视频序列方面取得了显著进展。与潜在空间世界模型和基于 Transformer 的世界模型不同，Sora 利用了一个编码网络将视频和图像转换为标记，然后通过大规模的扩散模型对这些标记进行加噪和去噪处理，随后将它们映射回原始图像空间，从而根据语言描述生成多步图像预测。这一能力使 Sora 在具体任务中具有高度适用性。例如，Sora 可以根据机器人任务描述和轨迹先验生成未来时间步的代理轨迹视频，从而增强基于模型的强化学习。UniPi [41] 使用扩散模型在图像空间中对代理轨迹进行建模，从语言输入和初始图像生成未来的关键视频帧，然后在时间序列中进行超分辨率以创建一致的高质量图像序列。UniSim [201] 通过联合训练基于互联网数据和机器人交互视频的扩散模型进一步改善了轨迹预测，能够预测高层次和低层次任务指令的长序列视频轨迹。

5.1.3 **联合嵌入预测架构** . 上述基于数据驱动的世界模型在自然语言处理任务中表现出色，但由于依赖于训练数据，缺乏真实世界的常识。由 Meta 的 Yann LeCun 提出的联合嵌入预测架构（Joint Embedding Predictive Architecture，JEPA）[92] 是一种突破性的方案，可以克服常识方面的局限性。受人类大脑高效学习的启发，JEPA 在高层表示空间中引入了分层规

划和自监督学习。分层规划将复杂任务分解为多个抽象层次，每个层次处理特定的子任务，以简化决策和控制，关注语义特征而非传统生成模型所关注的像素级输出。通过自监督学习，JEPA 训练网络去预测缺失或隐藏的输入数据，使其能够在大型未标记数据集上进行预训练并针对多样任务进行微调。JEPA 的架构包括一个感知模块和一个认知模块，利用潜变量形成一个世界模型，以捕捉关键信息，同时过滤冗余，支持高效的决策和未来情境规划。通过结合双系统概念，JEPA 在“快速”直观反应与“缓慢”深思熟虑的推理之间实现平衡。这种分层规划、自监督学习和强大世界模型的结合使 JEPA 成为一个适用于复杂真实环境的可扩展、认知启发的框架。

5.2 决策中的世界模型

世界模型可以为代理提供强大的内部表示，使其能够在实际行动之前预测环境动态和潜在结果。在决策过程中，世界模型发挥两个主要作用：(1) 模拟验证和 (2) 知识增强，如图 15 的左侧所示。通过这些机制，世界模型可以显著提高代理在复杂动态环境中规划和执行任务的能力。

5.2.1 **用于模拟验证的世界模型**. 在机器人领域，决策测试可能极其昂贵且耗时，尤其是对于那些顺序和长期任务而言，因为当前的决策会深刻影响未来的表现。世界模型可以通过实现模拟验证来缓解这一问题，使代理能够“尝试”执行动作并观察可能的后果，而无需做出实际的承诺。这种模拟验证大大缩短了迭代时间，并促进了在边缘情况或高风险场景下的安全测试，这些在现实中是不可行的。预测动作如何影响未来环境状态的能力帮助代理识别并避免潜在错误，最终优化性能。NeBula [2] 通过贝叶斯过滤构建概率信念空间，使机器人能够有效地在多种结构配置中进行推理，包括未知环境，提供了一种在不确定性下预测结果的复杂方法。UniSim [201] 是一个用于现实世界交互的生成模拟器，可以模拟高级指令和低级控制的视觉结果。它包含一个统一的生成框架，将动作作为输入，整合了不同调制下的多样化数据集。

5.2.2 **用于知识增强的世界模型**. 为了成功完成现实世界的任务，智能体通常需要丰富的知识和环境常识。世界模型可以通过提供预测和上下文知识来增强智能体，这对于策略规划至关重要。通过预测未来的环境状态或丰富智能体对世界的理解，世界模型使智能体能够预测结果、避免错误，并随着时间的推移优化性能。世界知识模型 (WKM) [135] 通过在任务开始前提供全局先验知识和在任务过程中维护本地动态知识，模仿人类的心理世界知识。它从专家和采样轨迹中合成全局任务知识和本地状态知识，与 LLM 集成时实现了卓越的规划性能。Agent-Pro [210] 将智能体与环境（尤其是与其他智能体在交互任务中的环境）互动的过程转化为“信念”。这些信念代表了智能体对环境的社会理解，并为随后的决策和行为策略更新提供信息。GovSim [133] 探索 LLM 智能体社会中合作行为的出现。这些智能体通过多智能体对话收集关于外部世界和其他智能体策略的信息，隐含形成其对世界模型的高层次洞察和表征。

除了决策之外，世界模型还可以使智能体高效地学习新技能和新行为。与通常因直接的智能体-环境交互而导致高计算成本和数据效率低下的无模型强化学习不同，基于模型的强化学习通过使用世界模型来简化学习过程，其方法是：(1) 模拟状态转换和 (2) 生成数据，如图 15 右侧所示。

5.2.3 **状态转换的世界模型**. 传统的强化学习是无模型的，直接通过代理与环境的交互进行学习，这种方式在数据稀缺或复杂场景中计算量大且不切实际。基于模型的强化学习通过利用能够明确捕捉状态转换和动态性的世界模型来缓解这些限制，使代理能够从模拟环境中提高其学习过程，实现安全、经济并高效的数据训练。世界模型创建现实世界的虚拟表示，以便代理可以探索假想的动作并优化策略，而无需承担与现实世界交互相关的潜在风险或成本。RobotDreamPolicy [134] 在世界模型中学习并发展策略，大幅减少与真实环境的交互。DayDreamer [191] 利用基于 RSSM 的 Dreamer V2 世界模型，将观察编码为潜状态并

预测未来状态，实现了高效样本学习中快速掌握机器人技能。SWIM [118] 更进一步，通过使用互联网规模的人类视频数据来理解丰富的人类互动并获得有意义的功能性启示。它最初在一个大型自我中心视频数据集上进行训练，然后通过机器人数据进行微调以适应机器人领域。随后，可在该世界模型中高效学习指定任务的行为。

除了增强学习和优化策略外，世界模型，特别是基于扩散的世界模型，还可以用于合成数据，这对于具身 AI 尤为有价值，因为收集多样且广泛的真实世界数据具有挑战性。基于扩散的世界模型可以合成逼真的轨迹数据、状态表示和动态性，从而扩充现有数据集或创建全新的数据集，以增强学习过程。SynthER [108] 利用基于扩散的世界模型生成低维度的离线强化学习轨迹数据，以增加原始数据集。他们的评估表明，扩散模型可以有效地从轨迹数据中学习状态表示和动态方程。MTDiff [69] 采用基于扩散的世界模型生成多任务轨迹，使用专家轨迹作为提示以指导生成与特定任务目标和动力学相符的代理轨迹。VPDD [68] 使用大规模的人类操作数据集训练轨迹预测世界模型，然后仅用少量标记的动作数据微调动作生成模块，从而显著减少政策学习所需的大量机器人交互数据。

6 挑战和未来前景

体现式智能呈现出了超越虚拟限制的前所未有的机会。然而，挑战仍然存在。在本节中，我们讨论高质量体现数据的稀缺、长期适应性的持续学习、计算和部署效率以及从仿真到现实的差距等主要未解决问题。通过研究核心挑战，我们指出了朝着构建健壮、适应性强且真正智能的体现系统的潜在研究方向。

6.1 具身数据的稀缺

培训具身智能体需要大量且多样的数据集。RT-X [175] 从 60 多个实验室收集了机器人手臂数据，并建立了开放的 X-Embodiment 数据集。AutoRT [4] 提出了在新环境中自动收集数据的系统。然而，真实世界的机器人数据仍然不足。原因在于机器人设计的巨大多样性、现实世界交互的复杂性以及各种任务的具体要求等。最先进的具身数据集，例如 VIMA [80]（具有 65 万次演示）和 RT-1 [26]（具有 13 万次演示），仍然相较于其视觉-语言对手，如 LAION-5B（具有 57.5 亿对文本-图像对），显得相形见绌。为了解决具身数据稀缺的问题，研究人员尝试了各种解决方案。

(1) 利用世界模型，特别是基于扩散的世界模型，从现有的智能体经验中合成新数据。SynthER [108] 利用基于扩散的世界模型来合成数据并增强离线 RL 轨迹数据集，从而显著提高在离线和在线环境中的性能。
(2) 整合大型人类数据集。Ego4D [53] 提供了从互联网视频中获取的丰富的现实世界动态和观察。这种方法通过利用常见的人类行为和互动来帮助提高机器人任务的上下文理解。然而，由于人类和机器人之间的形态差异，直接将人类动作转移到机器人通常会导致不对齐和转移性降低。

通过世界模型生成数据强调数据的质量和多样性，而人类数据整合则利用了真实世界的背景。但它们仍然面临现实差距、计算成本和对齐问题的挑战。

6.2 持续学习

具身智能系统应能够通过与开放动态环境的持续交互自主更新知识和优化策略，同时在不断变化的任务和条件下保持先前获得的能力。这种能力可以通过持续学习来实现 [117] 。如果没有持续学习，智能代理需要为每个新场景或轻微的环境变化重新训练，这严重限制了其在现实世界中的实用性。然而，持续学习存在重大障碍。为了解决这些挑战，研究人员正在探索各种方法。体验重放 [8] 可以通过定期回顾历史数据来缓解灾难性遗忘。正则化技术 [89] 通过在新任务学习期间限制权重更新来保留先前任务的知识。数据混合策略 [90] 将先前数据分布与新数据按不同比例结合，以减少特征失真。诸如 CycleResearcher [188] 之类的框架通过优化策略和奖励模型促进复杂过程中的具身学习。未来的进步可能包括增强自监

督学习，以通过内在动机驱动主动探索，并结合多智能体协作机制，通过集体交互加速个体学习。

日益复杂的具身智能模型在训练和部署时需要大量的计算资源。例如，DiffusionVLA [20] 需要数百个高端 GPU 以及百万规模的轨迹数据集上耗费数周的训练，计算量达到千万亿次浮点运算（PFLOPs）。其推理时的迭代采样导致了几秒钟的延迟，这对机器人实时控制应用是一种阻碍。基于 Transformer 的 VLA RT-2 [222] 保持了一个复杂的架构，要求大约 20GB 的视频内存。虽然 RT-2 通过预训练降低了训练成本，这种高内存需求却增加了在资源受限的边缘设备（如实际机器人）上部署的困难。基于云的部署作为一种替代方案通常不切实际，因为涉及到与物理机器人互动所固有的数据隐私、安全和实时操作限制问题。为缓解这些挑战，正在探索几种策略。通过压缩技术优化大规模模型和设计固有的轻量级架构是实现具身 AI 在边缘设备上有效且广泛部署的最可行的方法。

6.3 模拟到现实的差距

具身人工智能需要大量数据来训练代理。然而，采集现实世界中各种具身的这种数据是非常昂贵或不切实际的。模拟器通过使代理能够在大量且多样化的模拟数据集 [152] 上进行训练来缓解这个问题，这被证明是一种具有成本效益且可扩展的解决方案。在模拟器中训练之后，代理通过模拟到现实的转移被部署到现实世界环境中。

然而，由于模拟环境与真实世界环境之间存在根本性的差异，仿真到真实的迁移面临“仿真到真实的差距”[176] 。这些差异以各种形式表现出来，例如不准确的物理动态 [30] 和视觉渲染的差异 [12] 。例如，摩擦、碰撞和流体行为难以精确建模；光照、相机曝光和材料特性难以模拟。在模拟环境中训练的代理在面对现实世界的细微不完善和复杂性时往往会意外失败，因为仿真无法完全复制现实。因此，训练良好的策略可能会在现实世界的分布外场景中失效。此外，精确建模现实世界环境本身就是一项艰巨的任务 [36] 。模拟和真实世界之间的细微差别往往会在长期决策中累积，导致策略不够稳健或无法适应环境变化。

先进的模拟器，例如可微分且高度逼真的 Genesis [124] ，正通过精确的物理建模和照片级逼真的渲染积极缩小这一差距，从而提高代理从模拟器到现实世界的可转移性。然而，弥合模拟到现实的差距仍然是实现稳健的具身人工智能的一个重大挑战。

7 结论

大模型的出现赋予了具身智能体强大的智能能力。本文对具身人工智能的技术和最新进展进行了全面的综述，重点是由大模型提供支持的自主决策和具身学习。我们首先通过介绍具身人工智能的基本知识和各种主要大模型开始这次综述，回顾它们在具身智能领域的最新发展和应用。然后，我们详细阐述了具身人工智能的决策方法，详尽介绍了分层和端到端范式、其基本机制以及最新进展。之后，我们回顾了具身学习机制，重点介绍模仿学习和强化学习，特别是大模型如何赋能它们。随后，我们介绍世界模型，展示其设计方法及其在决策和具身学习中的重要作用。最后，我们讨论了具身智能中的开放性挑战，包括具身数据稀缺、持续学习、计算和部署效率以及从模拟到现实的差距，并提出潜在的解决方案。通过这次系统的调查研究，我们为研究人员和工程师提供了关于具身人工智能领域现状和开放性挑战的深入总结与分析，同时指出了通向人工通用智能的潜在发展方向。

References

[1] Abdul Afram and Farrokh Janabi-Sharifi. 2014. Theory and applications of HVAC control systems–A review of model predictive control (MPC). *Building and environment* 72 (2014), 343–355.

[2] Ali Agha, Kyohei Otsu, Benjamin Morrell, David D Fan, Rohan Thakker, Angel Santamaria-Navarro, Sung-Kyun Kim, Amanda Bouman, Xianmei Lei, Jeffrey Edlund, et al. 2021. Nebula: Quest for robotic autonomy in challenging environments; team costar at the darpa subterranean challenge. *arXiv preprint arXiv:2103.11470* (2021).

[3] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[4] Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, et al. 2024. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963* (2024).
[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
[6] Shan An, Ziyu Meng, Chao Tang, Yuning Zhou, Tengyu Liu, Fangqiang Ding, Shufang Zhang, Yao Mu, Ran Song, Wei Zhang, et al. 2025. Dexterous manipulation through imitation learning: A survey. *arXiv preprint arXiv:2504.03515* (2025).
[7] Daman Arora and Subbarao Kambhampati. 2023. Learning and leveraging verifiers to improve planning capabilities of pre-trained language models. *arXiv preprint arXiv:2305.17077* (2023).
[8] Benedikt Bagus and Alexander Gepperth. 2021. An investigation of replay-based approaches for continual learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.
[9] Andrew G Barto. 2021. Reinforcement learning: An introduction. by richard' s sutton. *SIAM Rev* 6, 2 (2021), 423.
[10] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. 2023. Hydra: Hybrid robot actions for imitation learning. In *Conference on Robot Learning*. PMLR, 2113–2133.
[11] Guillaume Bellegarda, Yiyu Chen, Zhuochen Liu, and Quan Nguyen. 2022. Robust high-speed running for quadruped robots via deep reinforcement learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10364–10370.
[12] Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems* 36 (2023), 35522–35543.
[13] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 17682–17690.
[14] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.
[15] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. 2024. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4788–4795.
[16] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164* (2024).
[17] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauzá, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. 2023. Robocat: A self-improving generalist agent for robotic manipulation. *arXiv preprint arXiv:2306.11706* (2023).
[18] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817* (2022).
[19] Rodney Brooks. 2003. A robust layered control system for a mobile robot. *IEEE journal on robotics and automation* 2, 1 (2003), 14–23.
[20] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. *OpenAI Blog* 1, 8 (2024), 1.
[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[22] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.
[23] Paweł Budzianowski, Wesley Maa, Matthew Freed, Jingxiang Mo, Winston Hsiao, Aaron Xie, Tomasz Młoduchowski, Viraj Tipnis, and Benjamin Bolte. 2025. EdgeVLA: Efficient Vision-Language-Action Models. *arXiv preprint arXiv:2507.14049* (2025).
[24] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. 2024. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems* (2024).
[25] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference*

on computer vision. 9650–9660.

[26] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3676–3713.

[27] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[28] Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, et al. 2023. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*. PMLR, 3909–3928.

[29] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[30] Po-Lin Chen and Cheng-Shang Chang. 2023. Interact: Exploring the potentials of chatgpt as a cooperative agent. *arXiv preprint arXiv:2308.01552* (2023).

[31] Shang-Fu Chen, Hsiang-Chun Wang, Ming-Hao Hsu, Chun-Mao Lai, and Shao-Hua Sun. 2023. Diffusion model-augmented behavioral cloning. *arXiv preprint arXiv:2302.13335* (2023).

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[33] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* (2023), 02783649241273668.

[34] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, et al. 2024. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775* (2024).

[35] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 45, 9 (2023), 10850–10869.

[36] Sahith Dambekodi, Spencer Frazier, Prithviraj Ammanabrolu, and Mark O Riedl. 2020. Playing text-based games with common sense. *arXiv preprint arXiv:2012.02757* (2020).

[37] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. 2024. Understanding world or predicting future? a comprehensive survey of world models. *Comput. Surveys* (2024).

[38] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. 2024. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812* (2024).

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[40] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).

[41] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. 2023. Learning universal policies via text-guided video generation. *Advances in neural information processing systems* 36 (2023), 9156–9172.

[42] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence* 6, 2 (2022), 230–244.

[43] Jonas Eschmann. 2021. Reward function design in reinforcement learning. *Reinforcement learning algorithms: Analysis and Applications* (2021), 25–33.

[44] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

[45] Rasool Fakoor, Pratik Chaudhari, Stefano Soatto, and Alexander J Smola. 2019. Meta-q-learning. *arXiv preprint arXiv:1910.00125* (2019).

[46] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. 2022. Implicit behavioral cloning. In *Conference on robot learning*. PMLR, 158–168.

[47] Maria Fox and Derek Long. 2003. PDDL2. 1: An extension to PDDL for expressing temporal planning domains. *Journal of artificial intelligence research* 20 (2003), 61–124.

[48] Gene F Franklin, J David Powell, Abbas Emami-Naeini, and J David Powell. 2002. *Feedback control of dynamic systems*. Vol. 4. Prentice hall Upper Saddle River.

[49] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219* (2020).
[50] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. 2024. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117* (2024).
[51] Michael Gelfond and Yulia Kahl. 2014. *Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach.* Cambridge University Press.
[52] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. 2023. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*. PMLR, 694–710.
[53] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.
[54] Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, et al. 2023. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. *arXiv preprint arXiv:2311.01977* (2023).
[55] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. 2023. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897* (2023).
[56] Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems* 36 (2023), 79081–79094.
[57] Yanjiang Guo, Yen-Jen Wang, Lihan Zha, and Jianyu Chen. 2024. Doremi: Grounding language model by detecting and recovering from plan-execution misalignment. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 12124–12131.
[58] Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine. 2018. Meta-reinforcement learning of structured exploration strategies. *Advances in neural information processing systems* 31 (2018).
[59] David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems* 31 (2018).
[60] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. Pmlr, 1861–1870.
[61] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603* (2019).
[62] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603* (2019).
[63] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2019. Learning latent dynamics for planning from pixels. In *International conference on machine learning*. PMLR, 2555–2565.
[64] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. 2020. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193* (2020).
[65] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104* (2023).
[66] Asher J Hancock, Allen Z Ren, and Anirudha Majumdar. 2024. Run-time observation interventions make vision-language-action models more visually robust. *arXiv preprint arXiv:2410.01971* (2024).
[67] Patrik Haslum, Nir Lipovetzky, Daniele Magazzeni, Christian Muise, Ronald Brachman, Francesca Rossi, and Peter Stone. 2019. *An introduction to the planning domain definition language*. Vol. 13. Springer.
[68] Haoran He, Chenjia Bai, Ling Pan, Weinan Zhang, Bin Zhao, and Xuelong Li. 2024. Large-scale actionless video pre-training via discrete diffusion for efficient policy learning. *CoRR* (2024).
[69] Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xuelong Li. 2023. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. *Advances in neural information processing systems* 36 (2023), 64896–64917.
[70] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.
[71] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems* 29 (2016).
[72] Kazuki Hori, Kanata Suzuki, and Tetsuya Ogata. 2024. Interactively robot action planning with uncertainty analysis and active questioning by large language model. In *2024 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 85–91.

[73] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278* (2025).
[74] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. 2023. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842* (2023).
[75] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. 2023. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176* (2023).
[76] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*. PMLR, 9118–9147.
[77] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973* (2023).
[78] Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, et al. 2023. Grounded decoding: Guiding text generation with grounded models for embodied agents. *Advances in Neural Information Processing Systems* 36 (2023), 59636–59661.
[79] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608* (2022).
[80] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094* 2, 3 (2022), 6.
[81] Yuqian Jiang, Harel Yedidsion, Shiqi Zhang, Guni Sharon, and Peter Stone. 2019. Multi-robot planning with conflicts and synergies. *Autonomous Robots* 43, 8 (2019), 2011–2032.
[82] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. 2023. Efficient diffusion policies for offline reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 67195–67212.
[83] Moo Jin Kim, Chelsea Finn, and Percy Liang. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645* (2025).
[84] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024).
[85] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024).
[86] Yeseung Kim, Dohyun Kim, Jieun Choi, Jisang Park, Nayoung Oh, and Daehyung Park. 2024. A survey on integration of large language models with intelligent robots. *Intelligent Service Robotics* 17, 5 (2024), 1091–1107.
[87] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.
[88] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.
[89] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.
[90] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054* (2022).
[91] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*. Springer, 45–73.
[92] Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* 62, 1 (2022), 1–62.
[93] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
[94] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision* 16, 1-2 (2024), 1–214.
[95] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. 2025. Pointvla: Injecting the 3d world into vision-language-action models. *arXiv preprint arXiv:2503.07511* (2025).

[96] Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. 2024. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16426–16435.
[97] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
[98] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
[99] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).
[100] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. 2022. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems* 35 (2022), 31199–31212.
[101] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. 2023. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378* (2023).
[102] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753* (2022).
[103] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*. 405–409.
[104] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. 2023. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots* 47, 8 (2023), 1345–1365.
[105] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
[106] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477* (2023).
[107] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics* (2025).
[108] Cong Lu, Philip Ball, Yee Whye Teh, and Jack Parker-Holder. 2023. Synthetic experience replay. *Advances in Neural Information Processing Systems* 36 (2023), 46323–46344.
[109] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093* (2024).
[110] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931* (2023).
[111] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2023), 46534–46594.
[112] Xinji Mai, Zeng Tao, Junxiong Lin, Haoran Wang, Yang Chang, Yanlan Kang, Yan Wang, and Wenqiang Zhang. 2024. From efficient multimodal models to world models: A survey. *arXiv preprint arXiv:2407.00118* (2024).
[113] Zhao Mandi, Shreeya Jain, and Shuran Song. 2024. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 286–299.
[114] Spyros Maniatopoulos, Philipp Schillinger, Vitchyr Pong, David C Conner, and Hadas Kress-Gazit. 2016. Reactive high-level behavior synthesis for an atlas humanoid robot. In *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 4192–4199.
[115] David Q Mayne, James B Rawlings, Christopher V Rao, and Pierre OM Scokaert. 2000. Constrained model predictive control: Stability and optimality. *Automatica* 36, 6 (2000), 789–814.
[116] Drew McDermott, Malik Ghallab, Adele E. Howe, Craig A. Knoblock, Ashwin Ram, Manuela M. Veloso, Daniel S. Weld, and David E. Wilkins. 1998. PDDL-the planning domain definition language. https://api.semanticscholar.org/CorpusID:59656859
[117] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. 2023. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research* 24, 214 (2023), 1–50.
[118] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. 2023. Structured world models from human videos. *arXiv preprint arXiv:2308.10901* (2023).

[119] Vincent Micheli, Eloi Alonso, and François Fleuret. 2022. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588* (2022).
[120] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.
[121] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems* 36 (2023), 25081–25094.
[122] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems* 36 (2023), 25081–25094.
[123] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.
[124] Rhys Newbury, Jack Collins, Kerry He, Jiahe Pan, Ingmar Posner, David Howard, and Akansel Cosgun. 2024. A review of differentiable simulators. *IEEE Access* (2024).
[125] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.
[126] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
[127] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
[128] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
[129] Chaofan Pan, Xin Yang, Yanhua Li, Wei Wei, Tianrui Li, Bo An, and Jiye Liang. 2025. A Survey of Continual Reinforcement Learning. *arXiv preprint arXiv:2506.21872* (2025).
[130] Cheng Pan, Xi Yang, Haoran Wang, Jiaming Zhang, Jushun Li, and Jie Xu. 2024. Hierarchical Continual Reinforcement Learning via Large Language Model. *arXiv preprint arXiv:2401.15098* (2024).
[131] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. 2023. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677* (2023).
[132] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747* (2025).
[133] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems* 37 (2024), 111715–111759.
[134] AJ Piergiovanni, Alan Wu, and Michael S Ryoo. 2019. Learning real-world robot policies by dreaming. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7680–7687.
[135] Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024. Agent planning with world knowledge model. *Advances in Neural Information Processing Systems* 37 (2024), 114843–114871.
[136] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. 2025. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830* (2025).
[137] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.
[138] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
[139] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[140] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access* 12 (2024), 26839–26874.

[141] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. 2019. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*. PMLR, 5331–5340.
[142] Sarthak S Raman, Vashisht Cohen, Erez Rosen, Shreyas Mirchandani, Benjamin Arkin, David Hedges, Alex Sorokin, Brent Gold, Tsung-Yen Fu, David Salac, et al. 2022. Planning with large language models via corrective re-prompting. *arXiv preprint arXiv:2210.03952* (2022).
[143] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
[144] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.
[145] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).
[146] Sonia Raychaudhuri and Angel X Chang. 2025. Semantic Mapping in Indoor Embodied AI–A Comprehensive Survey and Future Directions. *arXiv preprint arXiv:2501.05750* (2025).
[147] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).
[148] Machel Reid, Yutaro Yamada, and Shixiang Shane Gu. 2022. Can wikipedia help offline reinforcement learning? *arXiv preprint arXiv:2201.12122* (2022).
[149] Brian Reily, Peng Gao, Fei Han, Hua Wang, and Hao Zhang. 2022. Real-time recognition of team behaviors by multisensory graph-embedded robot learning. *The International Journal of Robotics Research* 41, 8 (2022), 798–811.
[150] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928* (2023).
[151] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. 2023. Transformer-based world models are happy with 100k interactions. *arXiv preprint arXiv:2303.07109* (2023).
[152] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. 2022. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on robot learning*. PMLR, 91–100.
[153] Gavin A Rummery and Mahesan Niranjan. 1994. *On-line Q-learning using connectionist systems*. Vol. 37. University of Cambridge, Department of Engineering Cambridge, UK.
[154] Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. 2025. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769* (2025).
[155] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
[156] Pratyusha Sharma, Antonio Torralba, and Jacob Andreas. 2021. Skill induction and planning with latent language. *arXiv preprint arXiv:2110.01517* (2021).
[157] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. 2024. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv:2403.12910* (2024).
[158] Ruizhe Shi, Yuyao Liu, Yanjie Ze, Simon S Du, and Huazhe Xu. 2023. Unleashing the power of pre-trained language models for offline reinforcement learning. *arXiv preprint arXiv:2310.20587* (2023).
[159] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 8634–8652.
[160] Keisuke Shirai, Cristian C Beltran-Hernandez, Masashi Hamaya, Atsushi Hashimoto, Shohei Tanaka, Kento Kawaharazuka, Kazutoshi Tanaka, Yoshitaka Ushiku, and Shinsuke Mori. 2024. Vision-language interpreter for robot task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2051–2058.
[161] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*. PMLR, 894–906.
[162] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
[163] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
[164] Tom Silver, Varun Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2022. PDDL planning with pretrained large language models. In *NeurIPS 2022 foundation models for decision*

making workshop.

[165] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2022. Progprompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302* (2022).

[166] Fuchun Sun, Runfa Chen, Tianying Ji, Yu Luo, Huaidong Zhou, and Huaping Liu. 2024. A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives. *CAAI Artificial Intelligence Research* 3 (2024).

[167] Ed D Tate, Jessy W Grizzle, and Huei Peng. 2009. SP-SDP for fuel consumption and tailpipe emissions minimization in an EVT hybrid. *IEEE Transactions on Control Systems Technology* 18, 3 (2009), 673–687.

[168] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[169] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213* (2024).

[170] Andrea Tirinzoni, Andrea Sessa, Matteo Pirotta, and Marcello Restelli. 2018. Importance weighted transfer of samples in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4936–4945.

[171] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[172] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[173] A. M. Turing. 1950. Computing machinery and intelligence. *Mind* LIX, 236 (1950), 433–460.

[174] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

[175] Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, et al. 2023. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*.

[176] Andrew Wagenmaker, Kevin Huang, Liyiming Ke, Kevin Jamieson, and Abhishek Gupta. 2024. Overcoming the Sim-to-Real Gap: Leveraging Simulation to Learn to Explore for Real-World RL. *Advances in Neural Information Processing Systems* 37 (2024), 78715–78765.

[177] Zishen Wan, Yuhang Du, Mohamed Ibrahim, Yang Zhao, Tushar Krishna, and Arijit Raychowdhury. 2024. Thinking and moving: An efficient computing approach for integrated task and motion planning in cooperative embodied ai systems. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*. 1–7.

[178] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).

[179] Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. 2025. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence* 4, 1 (2025), 52–64.

[180] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805* (2024).

[181] Ziyi Wang, Sheng Cai, Guandao Chen, Si-Yuan Chen, Hong-Xin Yang, Shuyuan Liu, Zihao Zhao, Yuxiang Wang, Chen Chen, Leo J Li, et al. 2023. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560* (2023).

[182] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. 2022. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193* (2022).

[183] Christopher John Cornish Hellaby Watkins et al. 1989. Learning from delayed rewards. (1989).

[184] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[185] Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, et al. 2024. Diffusion-VLA: Generalizable and Interpretable Robot Foundation Model via Self-Generated Reasoning. *arXiv preprint arXiv:2412.03293* (2024).

[186] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. 2025. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855* (2025).
[187] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, et al. 2025. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters* (2025).
[188] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816* (2024).
[189] David E Wilkins. 2014. *Practical planning: extending the classical AI planning paradigm*. Elsevier.
[190] Lik Hang Kenny Wong, Xueyang Kang, Kaixin Bai, and Jianwei Zhang. 2025. A Survey of Robotic Navigation and Manipulation with Physics Simulators in the Era of Embodied AI. *arXiv preprint arXiv:2505.01458* (2025).
[191] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. 2023. Daydreamer: World models for physical robot learning. In *Conference on robot learning*. PMLR, 2226–2240.
[192] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. 2023. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848* (2023).
[193] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Shuo Jiang, Bin He, and Qian Cheng. 2025. Robot learning in the era of foundation models: A survey. *Neurocomputing* (2025), 129963.
[194] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. 2024. Text2reward: Automated dense reward function generation for reinforcement learning. In *International Conference on Learning Representations (ICLR), 2024 (07/05/2024-11/05/2024, Vienna, Austria)*.
[195] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. 2020. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908* (2020).
[196] Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. 2022. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*. PMLR, 24631–24645.
[197] Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. 2025. Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation. *arXiv preprint arXiv:2502.02175* (2025).
[198] Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. 2024. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385* (2024).
[199] Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. 2024. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385* (2024).
[200] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Haoran Tan, Chencheng Jiang, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. 2024. Octopus: Embodied vision-language programmer from environmental feedback. In *European conference on computer vision*. Springer, 20–38.
[201] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. 2023. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114* 1, 2 (2023), 6.
[202] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.
[203] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
[204] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. 2023. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647* (2023).
[205] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access* 8 (2020), 58443–58469.
[206] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 2024. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954* (2024).
[207] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* (2023).
[208] Rongyu Zhang, Menghang Dong, Yuan Zhang, Liang Heng, Xiaowei Chi, Gaole Dai, Li Du, Yuan Du, and Shanghang Zhang. 2025. Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. *arXiv preprint arXiv:2503.20384* (2025).
[209] Tianyao Zhang, Xiaoguang Hu, Jin Xiao, and Guofeng Zhang. 2022. A survey of visual navigation: From geometry to embodied AI. *Engineering Applications of Artificial Intelligence* 114 (2022), 105036.
[210] Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. 2024. Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint*

arXiv:2402.17574 (2024).

[211] Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. 2023. Storm: Efficient stochastic transformer based world models for reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 27147–27166.

[212] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361* (2023).

[213] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705* (2023).

[214] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023).

[215] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631* (2024).

[216] Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. 2025. Lifelong learning of large language model based agents: A roadmap. *arXiv preprint arXiv:2501.07278* (2025).

[217] Qinqing Zheng, Amy Zhang, and Aditya Grover. 2022. Online decision transformer. In *international conference on machine learning*. PMLR, 27042–27059.

[218] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. 2024. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345* (2024).

[219] Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Yitao Liang, Yaodong Yang, and Yuanpei Chen. 2025. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900* (2025).

[220] Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma. 2024. Isr-llm: Iterative self-refined large language model for long-horizon sequential task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2081–2088.

[221] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. 2024. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520* (2024).

[222] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*. PMLR, 2165–2183.