中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)06-2017-09

论文引用格式:Mu Y, Zhao H, Hu R Z, Zhang L, Li H Y, Yang J L, Wang J B, Han L, Su Y F, Xu K, Yang Y, Li J, Dai R L, Chen B Q, Liu Y B and Yi L. 2026. Frontiers and prospects of embodied AI: evolution of data, models, and systems. Journal of Image and Graphics, 31(6):2017-2025
(穆尧, 赵昊, 胡瑞珍, 张力, 李弘扬, 杨蛟龙, 王靖博, 韩磊, 苏永峰, 徐凯, 杨易, 李江, 戴若犁, 陈宝权, 刘烨斌, 弋力. 2026. 具身智能前沿展望:数据、模型与系统演进. 中国图象图形学报, 31(6):2017-2025)[DOI:10.11834/jig.260059]

具身智能前沿展望:数据、模型与系统演进

穆尧[1],赵昊[2],胡瑞珍[3],张力[4],李弘扬[5],杨蛟龙[6],王靖博[7],韩磊[8],苏永峰[9],徐凯[10],杨易[11],李江[9],戴若犁[8],陈宝权[12],刘烨斌[2],弋力[2*]

1. 上海交通大学计算机学院,上海　200240; 2. 清华大学智能产业研究院,北京　100084; 3. 深圳大学计算机软件学院,深圳　518055;
4. 复旦大学大数据学院,上海　200433; 5. 香港大学火枪手基金会数据科学研究院,香港　999077; 6. 微软亚洲研究院,北京　100080;
7. 上海人工智能实验室,上海　200444; 8. 诺亦腾机器人科技(深圳)有限公司,深圳　518048;
9. 深圳引望智能技术有限公司,深圳　518110; 10. 国防科技大学计算机学院,长沙　410073;
11. 浙江大学计算机科学与技术学院,杭州　310058; 12. 北京大学人工智能研究院,北京　100871

摘　要:具身智能作为人工智能发展的关键领域,正面临数据异构性、强物理约束及交互昂贵等挑战,难以直接复制大语言模型的"大规模预训练+规模定律"范式。本文从数据、模型、系统与评测4个维度全面梳理了具身智能的前沿技术演进。在数据层面,提出了"数据金字塔"结构,主张利用底层庞大的仿真与互联网视频数据构建物理常识,通过中层人类交互数据进行行为映射,最终以顶层少量真机数据实现技能落地;在模型层面,探讨了主流视觉—语言—动作模型(vision-language-action,VLA)的扩展瓶颈,并指出"世界模型"作为具身预训练的新方向,能够通过模拟环境动力学与未来预演,赋予智能体更强的物理直觉与泛化能力;在系统层面,观察到架构正从单一端到端模型向类操作系统的"分层架构"演进,实现高层语义规划与底层运动控制的解耦。最后,本文审视了当前评测体系在真实性与可复现性上的挑战,并对行走与操作一体化及具身智能"ImageNet时刻"的到来进行了展望。

关键词:具身智能;数据金字塔;世界模型;VLA模型;分层控制架构;具身评测

Frontiers and prospects of embodied AI: evolution of data, models, and systems

Mu Yao[1], Zhao Hao[2], Hu Ruizhen[3], Zhang Li[4], Li Hongyang[5], Yang Jiaolong[6], Wang Jingbo[7], Han Lei[8], Su Yongfeng[9], Xu Kai[10], Yang Yi[11], Li Jiang[9], Dai Ruoli[8], Chen Baoquan[12], Liu Yebin[2], Yi Li[2*]

1. School of Computer Science, Shanghai Jiao Tong University, Shanghai　200240, China; 2. Institute for Intelligent Industry, Tsinghua University, Beijing　100084, China; 3. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen　518055, China; 4. School of Data Science, Fudan University, Shanghai　200433, China; 5. Musketeers Foundation, Institute of Data Science, Hong Kong University, Hong Kong　999077, China; 6. Microsoft Research Asia, Beijing　100080, China; 7. Shanghai Artificial Intelligence Laboratory, Shanghai　200444, China; 8. Noitom Technology Co., Ltd., Shenzhen　518048, China; 9. Shenzhen Yinwang Intelligent Technology Co., Ltd., Shenzhen　518110, China; 10. College of Compute, National University of Defense Technology, Changsha　410073, China; 11. College of Computer Science and Technology, Zhejiang University, Hangzhou　310058, China; 12. Institute for Artificial Intelligence, Peking University, Beijing　100871, China

收稿日期:2026-01-28;**修回日期**:2026-04-07;**预印本日期**:2026-04-14

*通信作者:弋力　ericyi0124@gmail.com

基金项目:国家自然科学基金项目(62125107, 62325211, 62132021, 62376060);新一代人工智能国家科技重大专项(2025ZD0123004);宁波市科技计划项目(2025Z038)

Supported by: National Natural Science Foundation of China (62125107, 62325211, 62132021, 62376060); New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123004)

Abstract: As a critical, rapidly evolving domain within artificial intelligence (AI), embodied AI represents the convergence of computer vision, natural language processing, and robotics, and aims to create intelligent agents capable of perceiving, reasoning, and acting within the physical world. However, despite the transformative success of large language models (LLMs) in the digital realm, embodied AI faces unprecedented, multifaceted challenges that hinder the direct replication of the "large-scale pre-training plus scaling law" paradigm. These challenges include extreme data heterogeneity across different robot morphologies, strong physical constraints that demand safety and precision, and the prohibitively expensive interaction costs associated with collecting real-world robotic data. Consequently, simply scaling up model parameters without addressing these domain-specific hurdles has proven insufficient for achieving general-purpose robotic intelligence. This paper comprehensively reviews the frontier technical evolution of embodied AI and offers a systematic analysis across four critical dimensions, namely, data, models, systems, and evaluation, to chart a path toward more robust and generalized embodied agents. In terms of data, this paper proposes a "Data Pyramid" structure designed to maximize data efficiency and transferability. This hierarchical framework advocates for the foundational use of massive, low-cost simulation data and Internet-scale video datasets at the bottom layer to build broad physical commonsense and visual representations; the utilization of human interaction data (such as ego-centric videos and teleoperation logs) in the middle layer to facilitate behavioral mapping and intent understanding; and the strategic application of a small, high-quality amount of real-world robot data at the top layer for fine-tuning and final skill deployment, an approach bridging the reality gap. Regarding models, the paper critically discusses the current state of mainstream vision-language-action models and highlights that while they excel at semantic understanding, they encounter significant scaling bottlenecks in continuous control and fine-grained manipulation. To overcome this, this paper identifies "World Models" as a pivotal new direction for embodied pretraining. By learning to simulate environmental dynamics, predict future states, and understand causal relationships without explicit supervision, world models promise to endow agents with deeper physical intuition and superior generalization capabilities in unseen environments. In terms of systems, this paper observes a paradigm shift, where the architecture is evolving from monolithic, single end-to-end models toward an operating system-like "Hierarchical Architecture." This evolution achieves the necessary decoupling of high-level semantic planning——powered by the reasoning capabilities of LLMs——and low-level motion control, which ensures precise execution and hardware compliance. This modular approach not only improves system robustness but also facilitates easier debugging and component upgrades. Finally, the paper examines the critical issues within current evaluation systems and specifically focuses on the challenges of authenticity in simulation benchmarks and the lack of reproducibility in real-world experiments. The field suffers from fragmented metrics that fail to capture the complexity of open-world interaction. In conclusion, this paper provides a forward-looking perspective on the inevitable integration of locomotion and manipulation——moving beyond stationary arms to mobile manipulators——and anticipates the arrival of the "ImageNet moment" for embodied AI, where standardized datasets and benchmarks will catalyze a Cambrian explosion of robotic capabilities, a development ultimately bridging the gap between digital intelligence and physical reality.

Key words: embodied AI; data pyramid; world models; VLA models; hierarchical control architecture; embodied evaluation

0 引 言

具身智能正站在人工智能发展的关键路口。过去十余年,人工智能(artificial intelligence,AI)在计算机视觉与大语言模型领域的突破,主要依赖于"大规模数据预训练+规模定律(scaling law)"的成功范式。然而,当人们将目光转向具身智能——让智能体在物理世界中感知、理解并行动——这一范式的直接迁移却遭遇了数据异构性、物理约束强及交互昂贵等前所未有的挑战。因此,如何构建高效的具身数据形态、设计能够理解时空连续性的模型架构,以及打造适应复杂任务的具身系统,成为当前学术界与产业界共同探索的前沿课题。

本文旨在全面梳理具身智能当前的技术演进,重点围绕数据、模型、系统与评测四大维度展开深度剖析。在数据层面,"数据金字塔"成为重要的发展趋势,即具身系统不再单纯依赖昂贵的真机遥操数

据，而是通过底层的仿真与互联网视频构建物理常识与语义理解，利用中层的人类交互数据作为行为映射的桥梁，最终通过顶层的少量真机数据实现技能的落地。在模型层面，探讨了主流的视觉—语言—动作模型（vision-language-action，VLA）在规模扩展上遇到的瓶颈，并指出了“世界模型”作为具身预训练新出路的潜力。世界模型通过模拟环境动力学与预演未来，有望赋予机器人更深层的物理直觉与泛化能力。在系统层面，观察到具身架构正从单一的端到端模型向类计算机操作系统的“分层架构”演进。通过解耦高层语义规划与底层运动控制，未来的机器人系统将具备更强的跨本体迁移能力与长程任务执行力。最后，审视当前评测体系面临的真实性与可复现性挑战。通过对这些关键节点的探讨，期望能为通向具身通用人工智能的路径提供清晰的注脚与思考。

1　数据：从“把地球遥操一遍”到数据金字塔

大语言模型的成功建立在一个朴素的假设之上：用全世界的文本数据训练一个模型，便可实现通用泛化。然而，类比到具身智能，这意味着需要所有机器人在所有任务上遥操一遍，这显然是不可能完成的任务。问题的本质不在于数据总量不够，而在于尚未找到能够支撑具身预训练的有效数据形态。当前，一种分层的“数据金字塔”结构（如图 1 所示）正在形成，它重新定义了不同数据源在智能体学习中的定位。

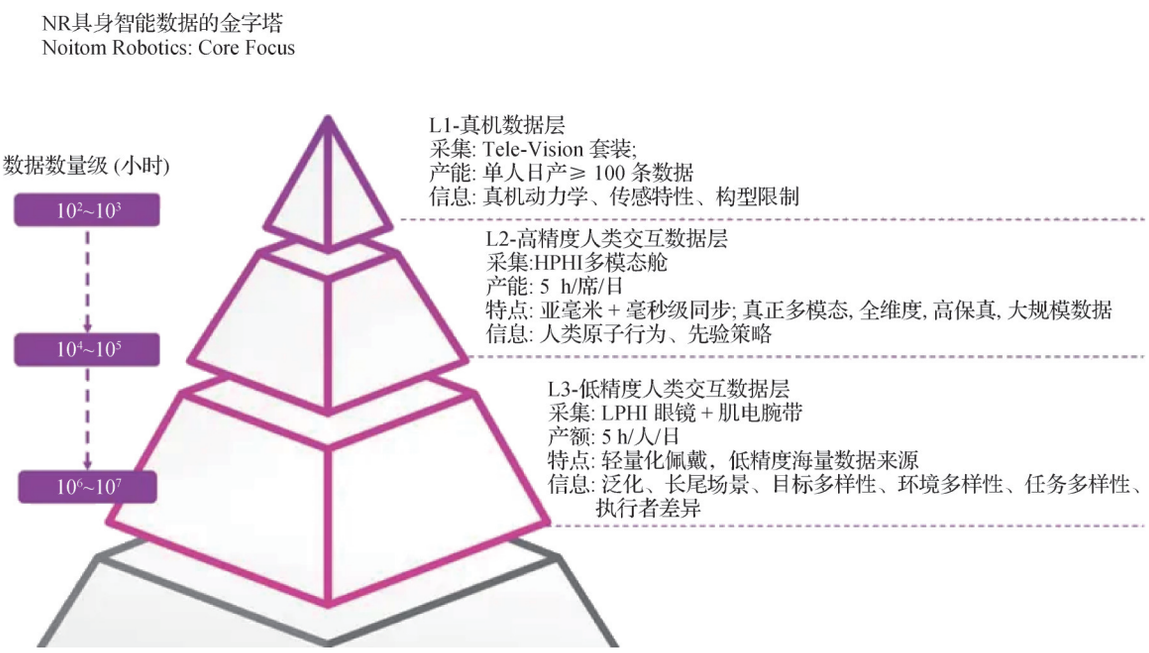

图 1　数据金字塔

Fig. 1　Data pyramid

1.1　基座层：互联网视频与仿真合成数据

互联网视频与仿真合成数据处于金字塔底层，规模最为庞大，包含极其丰富的物理常识、物体语义和长尾场景。

一方面，面对真实数据昂贵且危险的局限，仿真合成数据正在经历从“手工搭建”到“生成式演进”的范式变革。传统的仿真数据主要作为学习基础物理交互（如“松手物体会下落”、“如何保持平衡”）的低成本来源，而前沿的合成数据管线正通过与生成式 AI 的深度融合，解决“多样性”与“保真度”的双重瓶颈。例如，利用生成式 AI 将互联网海量数据转化为仿真资产（如 GenManip（Gao 等，2025）），构建出无穷变化的“生成式环境”，将互联网数据的语义多样性注入仿真训练；同时，建立由多模态大模型驱动的“生成—执行—评估—修正”闭环（如 RoboTwin（Chen 等，2025a），InternData-A1（Tian 等，2025）），打造无需人工干预的“自进化数据工厂”。这种方式能够批量产出兼具物理可行性与高层语义逻辑的专家级轨迹，极大地提升了仿真数据在复杂长程任务中的有效性，使其不再仅仅是真实数据的“平替”，而是

连接虚拟与现实的重要桥梁。

另一方面，互联网视频在场景、物体和技能等维度的多样性远超任何真机数据集，是一直以来被低估的数据金矿。微软亚洲研究院的VITRA（Li等，2025b）等研究表明，互联网上的人类视频蕴含了极丰富的场景语义和物体知识，通过将“人手”视为机器人末端，利用三维视觉技术构建数据管线，可将海量视频转化为结构化的VLA数据。基于此类数据预训练的模型，在未见过的真实家庭场景中展现出了惊人的零样本泛化能力——例如，模型能自发学会“抓瓶子时手指微张，捏铅笔时两指并拢”的通用操作逻辑。此外，仿真合成数据、互联网视频数据等，也在构筑世界模型、赋能具身推理方面展现出了巨大潜力，成为具身数据金字塔的基座。

1.2 桥梁层：人类交互数据

这一层包括第一视角人类交互数据与人体动捕数据，聚焦于解决机器人如何像人一样交互的问题，是连接抽象语义与具体控制的桥梁。首先，轻量级in-the-wild的第一视角人类交互数据，可突破实验室"笼子"，以轻便的AI眼镜、可穿戴手环（如Apple Vision Pro）等方式在真实世界中采集，达到中等数据规模，更贴近未来大模型所需的数据密度与多样性。美国加州大学圣迭戈分校王小龙团队的OpenTeleVision（Cheng等，2024）和Anyteleop（Qin等，2023）直接利用人本位数据进行训练，无需绑定特定真机，通过关键点对齐与姿态重定向，将末端信息映射到人形机器人上，实现human-centric与robot-centric数据的联合训练。其次，动捕环境中的高精度人—环境交互数据，可以在严格可控的环境中获得类真值级别的标注与多模态信号（包括力/触觉/视觉/听觉/高精度目标位置等）。相较真机遥操作数据，采集效率可提升1～2个数量级，因为摆脱了对具体真机的强绑定，还能系统化地做多目标与大空间覆盖。

1.3 顶层：真机遥操作数据

位于金字塔顶端，数据直接绑定机器人真机，包含真实的物理反馈与特定本体的动力学特征。优点是数据直接绑定真机末端，量化指标清晰，适合行为克隆或监督学习。但也存在天然约束：必须依赖真实本体，操作者受视角和交互方式限制，采集过程不够自然，效率偏低且成本极高。近年来，不少厂商和研究机构的做法是建设“机器人数据工厂”，如智元AgiBot World Colosseo（Team AgiBot-World，2025），以集中化、流水线的方式规模化采集。然而，采集环境往往被简化为桌面摆放若干物体的场景，与复杂的真实家庭等开放环境存在明显差距——这种差距不仅体现在环境布置与外观上，更包括物体多样性、任务复杂度、技能组合与层级结构的缺失。总体来看，遥操数据规模最小也最昂贵，往往用于最终的技能落地与“最后一公里”的精度打磨。

1.4 理想数据通路

基于上述金字塔结构，理想的数据使用路径并非简单的混合，而是一个从底向上的渐进学习过程。首先利用仿真合成数据与互联网视频进行大规模预训练。仿真数据让模型学习基础的物理交互（平衡感、接触感），互联网视频赋予模型对开放世界物体的语义理解能力。其次引入人类交互数据。通过重定向（retargeting）技术，将人类的灵巧操作映射到机器人空间。这一步不仅是为了学习完成任务，更是为了学习“合理的运动流形”，避免机器人出现怪异、危险的动作。最后使用真机遥操作数据进行微调，并持续从真机经验中学习。此时模型已具备常识与动作先验，只需极少量的真机数据即可将策略“坍缩”到具体的机器人本体上，实现从“通用智能”到“专用具身”的跨越，并在持续学习中不断完善自身能力，如图2所示。

在视觉主导的金字塔之外，触觉与力觉数据的缺失是当前具身智能迈向“精细操作”的最大隐痛。物理世界的本质不仅是“看”，更是“接触”。受限于当前触觉传感器（如GelSight、电子皮肤）的非标与高昂成本，相关数据极度稀缺。未来，构建包含力/触觉的大规模多模态数据集将是必经之路；只有补齐这一维度，模型才能真正理解摩擦力与柔性形变，实现从简单的“轨迹模仿”到具备真正物理交互智慧（如盲操作、防滑控制）的质变。

2 模型：VLA与世界模型的融合演进

2.1 VLA的困境：规模定律的缺席

视觉—语言—动作模型（VLA）是当前具身智能的主流路线，其基本思路是沿用视觉语言模型（vision-language model，VLM）的预训练方式，将连续动作离散化后进行下一个token预测。然而，一个值得注意的现象是从最早70亿参数的Open VLA，到

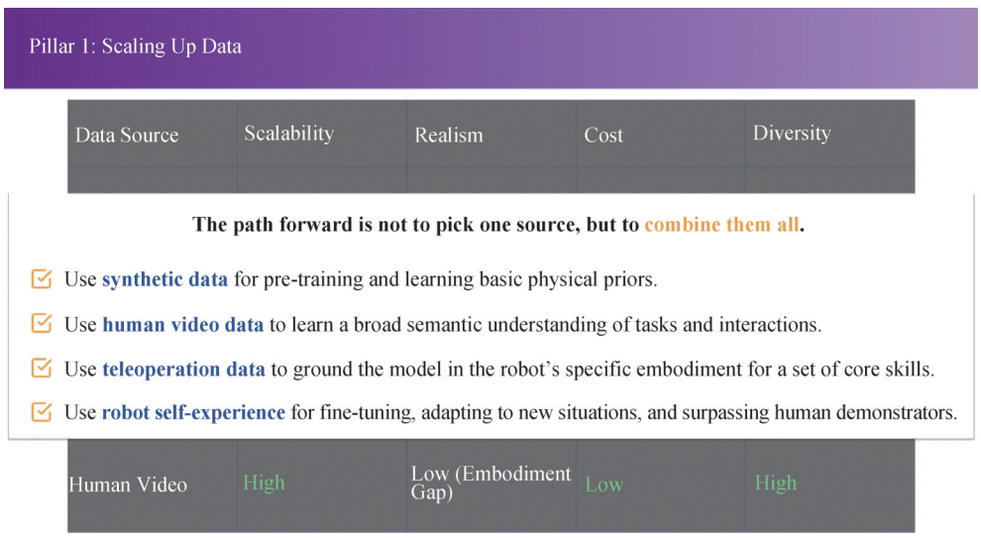

图 2　数据混合通路

Fig. 2　Hybrid data path

后来的 pi0（26 亿参数），再到 pi0. 5 及字节跳动、自变量等模型（均在 20~30 亿参数左右），由于大模型在小数据量训练易出现过拟合，参数规模不升反降，具身智能目前尚未找到属于自己的规模定律。

VLA 路线面临的核心困境在于：它严重依赖真机数据进行预训练，而真机数据存在规模不足、模态不统一以及本体差异大等问题，导致难以扩展。更本质的是，VLA 沿用 VLM 的思路进行序列预测，缺乏对物理世界时空连续性的深刻理解。为此，引入三维/四维时空先验（如 4D-VLA（Zhang 等，2025a）、DreamVLA（Zhang 等，2025c）、Spatial Forcing（Li 等，2025a））等改进措施，正成为增强 VLA 模型空间推理能力的重要方向。

2. 2　世界模型：具身训练的新出路

在此背景下，世界模型辅助具身学习作为一种新的技术路线正在崭露头角。具体而言，在具身智能中，世界模型可以视做一种通用“仿真器”：它能够根据当前的观测或状态，在不同的操控行为下预演未来。从控制论视角来看，世界模型可做狭义界定：对“被控系统”进行描述的动力学模型，其核心能力是“预测”——而预测正是规划与控制的前提。从 VPP（video prediction policy）（Hu 等，2025）到 Genie-Envisioner（Liao 等，2025），再到宇树开源的世界模型 UnifoLM-WMA-0（Team Unitree Robotics，2025），基本思路都是“视频生成模型 + 动作头”，整个领域似乎正在转向将世界模型与 VLA 相结合的路线，如图 3 所示。

世界模型相对传统 VLA 的核心优势在于：如果使用视频生成进行预训练，就没有任何扩展障碍，同时还能规避对硬件、本体和真机数据的依赖。需要特别指出的是，过去人们对世界模型的认知是“生成数据、生成环境，用来训练或评估另一个模型”。但对具身智能而言，这并非其最关键的价值。世界模型的核心价值在于：它可能是解决具身智能预训练问题的根本出路。在预训练阶段，需要的是积累对物理世界的“直觉认知”，而非追求完美的物理正确性——正如训练 LLM（large language model）时并不需要完美的答案一样，物理正确性可以在后训练（post-train）阶段再去完善。

更重要的是，世界模型在本质上是“任务无关”的——它关注的是环境动力学本身，能够仿真并“预演”在不同动作下任务的完成情况，而不仅是某个具体任务的最优动作分布。这一特性使其具备强大的可迁移性：一旦环境的动态规律被较好地捕捉，便可在多个任务之间复用，甚至服务于不同形态的智能体。基于模型的强化学习往往比完全无模型的方法具有更高的样本效率。通过在内部世界模型中进行“预演”，智能体可以在不与真实世界进行昂贵交互的情况下，对策略进行高效的评估和改进。这正是近期提出的 ProphRL（reinforcing action policies by prophesying）（Zhang 等，2025b）等工作的切入口：将预训练好的世界模型视做统一的“神经模拟环境”，在其中对 VLA 策略进行大规模后训练。类比 LLM 领域的 RLHF（reinforcement learning from human

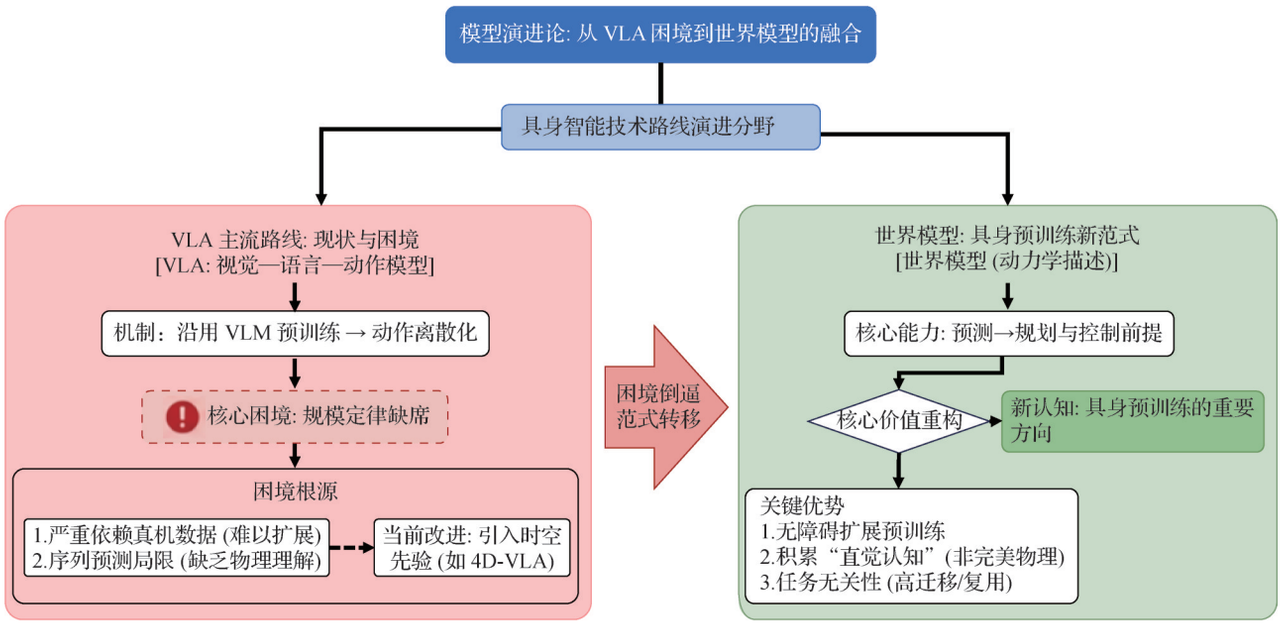

图3　VLA与世界模型的融合

Fig. 3　The integration of VLA and world models

feedback),这种方法通过奖励信号对策略进行再塑造,利用高保真视频生成环境中的大规模“想象轨迹”驱动策略更新,最终再将优化后的策略回迁到真实机器人上完成闭环,从而有效解决了真实环境交互昂贵且存在安全风险,以及传统物理仿真器难以兼顾多模态和复杂接触动力学的难题。

3　系统:分层控制架构

具身智能的落地不仅需要强大的模型(大脑),更需要合理的系统设计来支撑复杂的长程任务与高效的技能习得。当前的系统演进正呈现出从端到端向类计算机操作系统的“分层架构”演进的明显趋势,如RoboMemory(Lei等,2026)、RoboOS(Tan等,2025)。具身指令编译器借鉴了计算机体系结构,将系统分为清晰的层级:顶层是人类语言或高层任务描述,中间层是“具身指令集”,底层是具体的运动控制与硬件驱动。编译器的作用就是将高层的感知与意图“转译”为底层的关节、力矩和轨迹指令。这种架构实现了行为表示与机器人本体的解耦,便于跨本体迁移。与此同时,技能库(Skill Library)解决了长程任务规划的难题。模型不再需要从头学习每一个动作,而是学习一个可组合的技能集合(如抓取、推、走)。上层大模型(LLM/VLM)负责语义理解和任务拆解,输出高层的技能token;底层策略网络负责将token转化为具体的控制信号。这种分层控制策略(hierarchical control)既利用了大模型的推理能力,又保证了底层控制的实时性与稳定性。未来的机器人将拥有像智能手机一样的“操作系统”。开发者只需调用高层的技能API(如robot. pick_up(cup)),而无需关心底层的电机控制。这将极大地降低具身智能的应用开发门槛。

4　评测:多样性、真实性与可扩展性的三重挑战

4.1　评测范式的现状与困境

当前具身智能的评测方式主要有3类。第1类是基于仿真引擎驱动的评测,如SimplerEnv(Li等,2024)和Libero(Liu等,2023),优势在于成本低、门槛小,物体位姿与场景变量可进行充分而严格的随机化,但面临“仿真到真实”的落差——视觉层面的域差(纹理、光照、材质)与物理层面的偏差(接触、摩擦、顺应性)。第2类是基于真实机器人的评测,如RoboChallenge(Yakefu等,2025),最大优势在于“真”,但代价高昂:实验室能搭建的场景有限,维护与运行需要持续的人力物力,更关键的是可复现性问题突出。第3类是利用世界模型进行情景展开并引入多模态模型评分,如EWMBench(Yue等,2025),相对可复现且可扩展,但物理真实性仍偏弱,

模型幻觉带来的评测偏差不可忽视。

4.2 好评测的标准：公平性是底线

具身智能评测的底线要求只有一个：公平性，而公平性的本质是可复现性。没有可复现性，一切对比都失去意义。在这条底线之上，评测应当追求3个进阶目标：多样性（场景、物体、任务、行为的广覆盖）、真实性（视觉与物理的双重逼真）和可扩展性（平台的可配置与可重组），如图4所示。当前主流评测以成功率为第一指标，但这无法全面反映模型的真实能力。评测体系需要向多维度演进：任务完成度（对多子目标任务的细粒度评测）、多模态理解能力（跨任务、跨指令的准确率与鲁棒性）、轨迹质量（平滑性、冗余度、抓取稳定性、放置精确性），以及在“未见物体、未见场景、未见组合”条件下的零样本与少样本迁移性能。

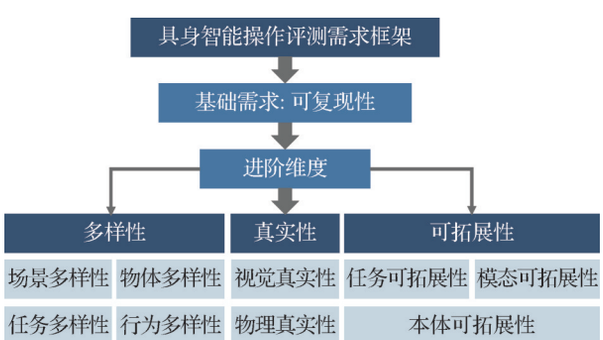

图4 具身评测的三重挑战

Fig. 4 The three challenges of embodied manipulation evaluation

4.3 走向标准化：双轨并行的建议

未来的评测范式建议“双轨并行、各司其职”。一方面，建设高度逼真的仿真环境，用有限而严格的仿真与真机场景支撑快速的算法迭代；另一方面，面向开放世界开展充分随机化的评测（指令、场景、光照、资产扰动等），考察模型的稳健性。

为维护真机评测的可复现性，需要推动标准化与平台化的举措。北美学术联盟提出的ManipulationNet（Chen等，2025b）提供了严格一致且可批量采购的真机资产清单，让社区能够在统一资产上公平对比；“Robot Arena”（Calife等，2009）式的平台则鼓励研究者提交策略，评测任务被分发到不同实验室以分布式方式统一执行。这些努力的共同指向，是将真机评测从“各做各的”引导到“统一口径、中心化协议、去中心化执行”。

5 展望：通向具身AGI的关键节点

5.1 Locomotion与Manipulation的一体化

如果说2024—2025年是将VLA更系统地引入具身智能的阶段，那么2026年很可能是“行走（locomotion）与操作（manipulation）一体化”规模推进的节点。将数据驱动的模仿学习范式从桌面级操作扩展到人形机器人的全身运动控制，有望催生出类似“Open-VLA”式的标志性全身控制模型。

在技术路径上，“离线预训练 + 在线微调”的组合范式正在成为主流：离线阶段采用监督/模仿学习建立稳定基座，在线阶段再通过微调、奖励学习与持续学习增强鲁棒性，并进一步缩小Sim-to-Real的差距。人形机器人的全身协调感知规划控制——包括双足行走、双臂操作、躯干平衡以及环境交互的统一建模——将成为这一阶段的核心技术挑战，而构建统一的世界模型，同时对机器人的全身本体状态与环境物体的视觉状态进行联合建模与预测，也成为有力的未来解决思路。

5.2 具身智能的“ImageNet时刻”何时到来

回顾AI发展史，每一次重大突破都伴随着标志性数据集或基准测试的出现：ImageNet之于计算机视觉，GLUE（general language understanding evaluation）/SuperGLUE之于自然语言理解。具身智能同样需要这样一个里程碑式的时刻。

李飞飞团队发起的BEHAVIOR Challenge（benchmark for everyday household activities in virtual, interactive, and real environments）（Li等，2023），可能成为具身智能的关键催化剂。这一挑战赛聚焦于家庭环境中的长程任务规划与执行，要求智能体在复杂、动态的场景中完成多步骤任务，预计将推动具身机器人算法的新一轮爆发。

然而，“ImageNet时刻”的真正到来，需要满足几个条件：1）数据层面，需要出现一个足够大规模、足够多样化且社区广泛认可的标准数据集；2）评测层面，需要建立公平、可复现且能区分算法优劣的基准测试；3）算法层面，需要有一个标志性的模型在该基准上取得突破性成绩，从而引发社区的广泛跟进。当这3个条件同时具备时，具身智能才可能迎来属于自己的“ImageNet时刻”，开启新一轮的快速发展。

参考文献(References)

Calife D, Bernardes J L Jr and Tori R. 2009. Robot Arena: an augmented reality platform for game development. Computers in Entertainment (CIE), 7(1): #11 [DOI: 10.1145/1486508.1486519]

Chen T X, Chen Z X, Chen B J, Cai Z J, Liu Y B, Li Z X, et al. 2025a. RoboTwin 2.0: a scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation [EB/OL]. [2026-01-07]. https://arxiv.org/pdf/2506.18088.pdf

Chen Y T, Kimble K, Adelson E, Asfour T and Chanrungmaneekul P. 2025b. ManipulationNet: benchmarking real-world robot manipulation at scale through physical skill challenges and embodied multimodal reasoning [CP/OL]. [2026-01-05]. https://manipulation-net.org/

Cheng X X, Li J L, Yang S Q, Yang G and Wang X L. 2024. Open-television: teleoperation with immersive active visual feedback// Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR 270: 2729-2749

Gao N, Chen Y L, Yang S, Chen X Y, Tian Y, Li H, et al. 2025. GenManip: LLM-driven simulation for generalizable instruction-following manipulation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 12187-12198 [DOI: 10.1109/CVPR52734.2025.01138]

Hu Y C, Guo Y J, Wang P C, Chen X Y, Wang Y J, Zhang J K, et al. 2025. Video prediction policy: a generalist robot policy with predictive visual representations//Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR 267:24328-24346

Lei M C, Cai H H, Yang Y Y, Wu Y M, Ren J K, Cui Z Z, et al. 2026. RoboMemory: a brain-inspired multi-memory agentic framework for interactive environmental learning in physical embodied systems [EB/OL]. [2026-01-07]. https://arxiv.org/pdf/2508.01415.pdf

Li C S, Zhang R H, Wong J, Gokmen C, Srivastava S, Martín-Martín R, et al. 2023. BEHAVIOR-1K: a benchmark for embodied AI with 1 000 everyday activities and realistic simulation//Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR: 80-93

Li F H, Song W X, Zhao H, Wang J B, Ding P X, Wang D L, et al. 2025a. Spatial forcing: implicit spatial representation alignment for vision-language-action model//Proceedings of 2025 International Conference on Learning Representations 2026 (ICLR). Rio de Janeiro, Brazil: #12276 [DOI: 10.48550/arXiv.2510.12276]

Li Q X, Deng Y, Liang Y B, Luo L, Zhou L, Yao C T, et al. 2025b. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos [EB/OL]. [2026-01-16]. https://arxiv.org/pdf/2510.21571.pdf

Li X L, Hsu K, Gu J Y, Mees O, Pertsch K, Walke H R, et al. 2024. Evaluating real-world robot manipulation policies in simulation// Proceedings of the 8th Conference on Robot Learning. Munich. Germany: PMLR: 3705-3728

Liao Y, Zhou P F, Huang S Y, Yang D L, Chen S C, Jiang Y X, et al. 2025. Genie envisioner: a unified world foundation platform for robotic manipulation//Proceedings of 2025 International Conference on Learning Representations 2026 (ICLR). Rio de Janeiro, Brazil: [s.n.].

Liu B, Zhu Y F, Gao C K, Feng Y H, Liu Q, Zhu Y K, et al. 2023. LIBERO: benchmarking knowledge transfer for lifelong robot learning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 44776-44791 [DOI:10.48550/arXiv.2306.03310]

Qin Y Z, Yang W, Huang B H, van Wyk K, Su H, Wang X L, et al. 2023. AnyTeleop: a general vision-based dexterous robot arm-hand teleoperation system//Proceedings of the Robotics: Science and Systems 2023. Daegu, Korea (South) [DOI: 10.15607/RSS.2023.XIX.015]

Tan H J, Hao X S, Chi C, Lin M L, Lyu Y X, Cao M Y, et al. 2025. RoboOS: a hierarchical embodied framework for cross-embodiment and multi-agent collaboration [EB/OL]. [2026-01-08]. https://arxiv.org/pdf/2505.03673.pdf

Team AgiBot-World. 2025. AgiBot world colosseo: a large-scale manipulation platform for scalable and intelligent embodied systems//Proceedings of 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE: 3549-3556 [DOI: 10.1109/IROS60139.2025.11247088]

Tian Y, Yang Y Y, Xie Y M, Cai Z T, Shi X, Gao N, et al. 2025. InternData-A1: pioneering high-fidelity synthetic data for pre-training generalist policy [EB/OL]. [2026-01-05]. https://arxiv.org/pdf/2511.16651.pdf

Team Unitree Robotics. 2025. UnifoLM-WMA-0: a world-model-action (WMA) framework under UnifoLM family[EB/OL].[2026-01-08]. https://unigen-x.github.io/unifolm-world-model-action.github.io

Yakefu A, Xie B, Xu C Y, Zhang E W, Zhou E J, Jia F, et al. 2025. RoboChallenge: large-scale real-robot evaluation of embodied policies [EB/OL]. [2026-01-08]. https://arxiv.org/pdf/2510.17950.pdf

Yue H, Huang S Y, Liao Y, Chen S C, Zhou P F, Chen L L, et al. 2025. EWMBench: evaluating scene, motion, and semantic quality in embodied world models//Proceedings of the 36th British Machine Vision Conference. Sheffield, UK: BMVA

Zhang J H, Chen Y R, Xu Y M, Huang Z, Zhou Y P, Yuan Y J, et al. 2025a. 4D-VLA: spatiotemporal vision-language-action pretraining with cross-scene calibration//Proceedings of the 39th Conference on Neural Information Processing Systems. San Diego, USA [DOI: 10.48550/arXiv.2506.22242]

Zhang J H, Huang Z, Gu C, Ma Z P and Zhang L. 2025b. Reinforcing action policies by prophesying [EB/OL]. [2026-01-08]. https://arxiv.org/pdf/2511.20633.pdf

Zhang W Y, Liu H S, Qi Z K, Wang Y N, Yu X Q, Zhang J Z, et al. 2025c. DreamVLA: a vision-language-action model dreamed with comprehensive world knowledge//Proceedings of the 39th Conference on Neural Information Processing Systems. San Diego, USA [DOI: 10.48550/ARXIV.2507.04447]

作者简介

穆尧，男，助理教授，主要研究方向为多模态具身智能和机器人学习。E-mail: muyao@sjtu.edu.cn

弋力，通信作者，男，助理教授，主要研究方向为具身智能和三维视觉。 E-mail: ericyi0124@gmail.com

赵昊，男，助理教授，主要研究方向为三维场景理解及其在机器人中的应用。E-mail: zhaohao@air.tsinghua.edu.cn

胡瑞珍，女，教授，主要研究方向为计算机图形学。E-mail: ruizhen.hu@szu.edu.cn

张力，男，教授，主要研究方向为计算机视觉与深度学习、世界模型、具身智能和自动驾驶。E-mail: lizhangfd@fudan.edu.cn

李弘扬，男，助理教授，主要研究方向为自动驾驶与具身智能。E-mail: hongyang@hku.hk

杨蛟龙，男，研究员，主要研究方向为计算机图形学和三维视觉。E-mail: jiaoyan@microsoft.com

王靖博，男，博士，主要研究方向为人形机器人学习。E-mail: wangjingbo@pjlab.org.cn

韩磊，男，高级工程师，主要研究方向为机器人和人工智能。E-mail: leihan.cs@gmail.com

苏永峰，男，工程师，主要研究方向为自动驾驶。E-mail: suyongfeng@yinwang.com

徐凯，男，教授，主要研究方向为计算机图形学、三维视觉及其机器人应用。E-mail: kevin.kai.xu@gmail.com

杨易，男，教授，主要研究方向为人工智能、计算机视觉和机器学习。E-mail: yangyics@zju.edu.cn

李江，男，工程师，主要研究方向为自动驾驶。E-mail: lijiang@yinwang.com

戴若犁，男，正高级工程师，主要研究方向为计算机视觉、多源传感器融合和机器人控制理论。E-mail: tristan@noitomrobotics.com

陈宝权，男，教授，主要研究方向为计算机图形学和三维视觉。E-mail: baoquan@pku.edu.cn

刘烨斌，男，教授，主要研究方向为三维视觉和计算成像。E-mail: liuyebin@tsinghua.edu.cn