基于世界模型的具身智能技术体系探索

宫丽娜1,2 徐嘉龙[1] 朱哲[1] 李安[1] 赵彦超!

(1. 南京航空航天大学计算机科学与技术学院，南京210000;
2. 南京航空航天大学深圳研究院，深圳518109)

摘 要： 文章通过深入分析具身智能技术的应用背景及现状，系统探讨了世界模型在提升具身智能决策泛化性与实时性中的核心作用，提出了“数据—模型—应用—评测”四层协同的世界模型助力具身智能技术体系框架，通过多模态数据治理、模型架构优化、任务场景应用与系统化评测实现闭环迭代与持续演进的技术路径。同时，以星动纪元的星动L7、极智嘉移动机器人等世界模型助力机器人为例，阐述了其在工业制造和物流领域关键任务场景中的实践效果，展示了世界模型从实验室走向规模化落地应用的示范意义。最后，文章展望物理可信生成、精准指令跟随、机器人导向评测与安全生成等重点突破方向，为具身智能技术产业化提供理论支撑与实践指引，推动其向安全、可靠、高效的方向持续发展。

DOI:10.12487/j.dt.2026.05.01

关键词： 具身智能；世界模型；物理仿真；规划与决策；多模态数据

中图分类号： TP18 **文献标志码：** A

Exploration of an Embodied Intelligence Technical System Based on World Models

Gong Lina[12]Xu Jialong'Zhu Zhe'LiAn'Zhao Yanchao[1]

(1.School of Computer Science and Technology/School of Artificial Intelligence,Nanjing University of Aeronautics and Astronautics,Nanjing 210000,China;
2.Shenzhen Research Institute of Nanjing University of Aeronautics and Astronautics,Shenzhen 518109,China)

Abstract: Through in-depth analysis of the application background and current status in the field of embodied intelligence,this paper systematically explores the core role of world models in enhancing the generalization capability and real-time performance of embodied intelligent decision-making,and proposes a four-layer collaborative technical framework of "data-model-application-evaluation"for world models to empower embodied intelligence,achieving a technical path of closed-loop iteration and continuous evolution through multimodal data governance,model architecture optimization,task-scenario applications,and systematic evaluation.At the same time,taking examples such as StarMove L7 from StarDynamics and Geek+mobile robots—machines empowered by world models—this paper elaborates on their practical performance in critical task scenarios across industrial manufacturing and logistics,demonstrating the exemplary significance of world models transitioning from laboratory research to large-scale real-world deployment.Finally,the paper looks forward to key breakthrough directions such as physically plausible generation,precise instruction following,robot-oriented evaluation, and safe generation,providing theoretical support and practical guidance for the industrialization of embodied intelligence technology,and promoting its sustained development toward safety,**reliability,**and efficiency.

Keywords: Embodied Intelligence;World Models;Physical Simulation;Planning and Decision-**Making;** Multimodal Data

随着新一轮科技革命和产业变革的加速演进，人工智能正从以感知理解为主的阶段迈向以自主决策与物理交互为核心的“行动智能”新范式。具身智能作为融合环境感知、情景认知、长时序推理与精准执行的新型智能形态，已成为全球科技竞争与未来产业布局的战略制高点。《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》《国务院关于深入实施“人工智能+”行动的意见》等纲领性文件，明确将具身智能及其核心使能技术列为重点突破方向，旨在推动高端制造、智慧物流等复杂场景实现柔性化、自主化与智能化升级，显著提升产业链韧性与自主可控水平。实践证明，具身智能不仅是培育新质生产力的关键载体，更是支撑“质量为先、效率优先、绿色低碳”国家战略落地的重要技术基石。

党的二十届四中全会进一步提出“加快高水平科技自立自强，引领发展新质生产力”“全面实施‘人工智能+’行动”I，为具身智能产业化指明方向。在从实验室验证向规模化商业落地的关键窗口期，具身智能与大模型、数字孪生、边缘计算等前沿技术的深度融合，正加速构建覆盖全生命周期的智能体化解决方案。然而，受真实交互数据稀缺、仿真—现实鸿沟、泛化能力不足及实时决策瓶颈等核心难题制约，行业与“干得更好、更灵活、更安全”的目标仍有明显差距。正是在此背景下，世界模型作为内化物理规律、预测环境动态并支持内部规划与决策的核心认知引擎，为破解上述瓶颈注入全新动能，成为推动具身智能从技术突破走向产业落地的决定性使能技术。

本文主要阐述了世界模型作为具身智能认知核心的关键作用，系统提出了“数据—模型—应用—评测”四层协同的技术体系框架。文章深入解析了从多模态数据治理、模型架构优化，到任务场景应用及系统化评测的闭环演进路径；结合工业制造与物流领域的落地案例，实证了世界模型在各方面的显著成效；最后，展望了物理可信生成、精准指令跟随、机器人导向评测与安全生成等重点突破方向，旨在为具身智能技术的产业化落地提供理论支撑与实践指引。

一、具身智能发展现状

2025年被业界公认为具身智能产业落地元年，我国相关企业已超150家2,部分头部企业实现小批量量产，商用服务机器人市场规模实现显著增长，家用、仓储、医疗等场景的试点应用加速拓展。东部地区依托上海、深圳、北京等创新集群优势，集聚优必选、宇树科技、智元机器人等领军企业，中西部地区则通过产业转移与专项基金加速追赶，整体形成全球领先的研发生态与产业链格局。欧美及日本等发达国家和地区依托特斯拉Optimus、Figure AI及波士顿动力等领军企业，在通用大模型与硬件协同上保持领先，聚焦于高保真仿真训练与真实场景数据闭环，加速推动人形机器人在汽车制造、物流分拣等高端领域的规模化验证，全球技术竞争态势日趋激烈。

然而，在产业的迅猛发展背后，技术与商业化落地仍面临显著挑战，呈现出“两极分化”态势。一方面，头部企业已初步构建“感知—规划—执行”技术栈，训练数据规模达较高水平，在结构化环境中任务完成率较高，并向真实场景初步迁移；另一方面，绝大多数中小初创企业仍局限于单模态模仿学习或特定任务，数据积累不足，跨场景迁移成功率较低，实时决策与物理交互稳定性不足。2025年行业融资额实现大幅增长，但真正形成稳定商业闭环的项目较少，技术成熟度与资本热度严重失衡。

具身智能技术实践主要体现出三个层面的特征：在研发训练阶段，依赖大规模仿真与视频

预训练使策略迭代周期显著缩短，但仅有少数企业建立高保真仿真环境库，真实世界数据稀缺导致模型过拟合严重；在具身执行层面，视觉—语言—动作（VLA）模型驱动的端到端方案使复杂操作成功率得到明显提升，头部企业在仓储分拣、家庭陪伴等场景开展试点，但全行业实时推理延迟仍较高，物理交互安全性与鲁棒性不足以支撑无监督大规模部署；在生态协同方面，开源社区初步形成，相似场景任务复用率处于中等水平，但数据隐私、标准缺失导致跨企业协作效率低下。这些矛盾集中反映出传统分层架构与数据驱动方法难以有效应对开放世界的动态性、长时序因果推理与多模态不确定性。因此，行业仍面临端到端模型泛化能力弱、“大脑—身体”协同优化难度大等问题，开放世界任务成功率普遍较低，高端复杂场景商业化落地率不足，“快而不稳、大而不精”特征突出。

正是在这一背景下，世界模型作为具身智能的认知核心，有效破解了传统机器人长期面临的三大核心痛点：场景泛化能力弱、长程任务规划困难以及实机试错成本过高。世界模型是理解现实世界动态的生成式神经网络。它以文本、图像、视频、运动等作为输入，模拟真实物理环境，并对运动、受力、空间关系等进行表示和预测。世界模型本质上是一种可学习的内部模拟系统。它通过编码器、动力学模型和解码器，构建了对外部物理世界的压缩表征与未来预测能力，从而使智能体能够在“脑内”拥有一个高度逼真的虚拟环境模拟器，而非仅仅依赖实时感知和反应式执行。与传统方法依赖固定规则的硬编码或纯数据驱动的模仿学习不同，世界模型赋予机器人类似于人类的“想象”与“预演”能力。它从海量视频、仿真数据或交互轨迹中学习物理规律、物体动态、因果关系以及环境转移模型，从而能够在给定当前观测和拟执行动作后，准确预测多步未来的状态、奖励，甚至不确定性。这种内部表征不仅能够捕捉静态物体属性，还能有效解耦动态要素，支持反事实推理和长时序预测。世界模型不再依赖对固定规则的死记硬背，而是通过构建可学习的内部物理引擎与因果推理机制，使机器人能够在“脑内”以极低成本进行虚拟推演：利用视频驱动的生成技术预演多种未来状态，结合扩散规划等算法在实际动作执行前优化决策路径，从而将原本高风险、高消耗的实体试错转变为高效的“思维实验”。这一从“走一步看一步”的反应式执行，向“先思考后行动”的预测式规划的范式跃升，不仅赋予机器人在未知环境中的举一反三能力与通用智能，而且大幅降低了规模化部署的技术门槛与安全代价，最终推动具身智能突破实验室演示的局限，实现从单体智能向群体协同、从零星试点向万台级规模部署的历史性跨越。

二、世界模型赋能具身智能的框架设计

在具身智能向能够适应和理解物理世界的通用智能演进的过程中，世界模型正逐渐从一种辅助性工具，演变为驱动智能体认知与决策的核心引擎。其角色已超越了传统意义上对环境动态的被动模拟，转而赋予智能体主动预测未来状态、规划行动序列、验证决策策略乃至理解底层物理规律的关键能力。为了系统性地实现这一范式转变，本文提出一套由四个层级构成的赋能框架(图1)。该框架以数据层为根基，整合来自真实与仿真环境的多模态感知数据，构建具身交互的认知基础；模型层作为核心，致力于构建兼具高生成保真度、动态可微性与深层语义理解能力的世界模型架构；应用层将世界模型的推演与规划能力转化为机器人在开放、动态场景中的具体感知、决策与控制功能；而评测层则贯穿始终，通过建立系统的评估体系，确保技术的发展始终锚定物理真实性、任务有效性与系统安全性三大

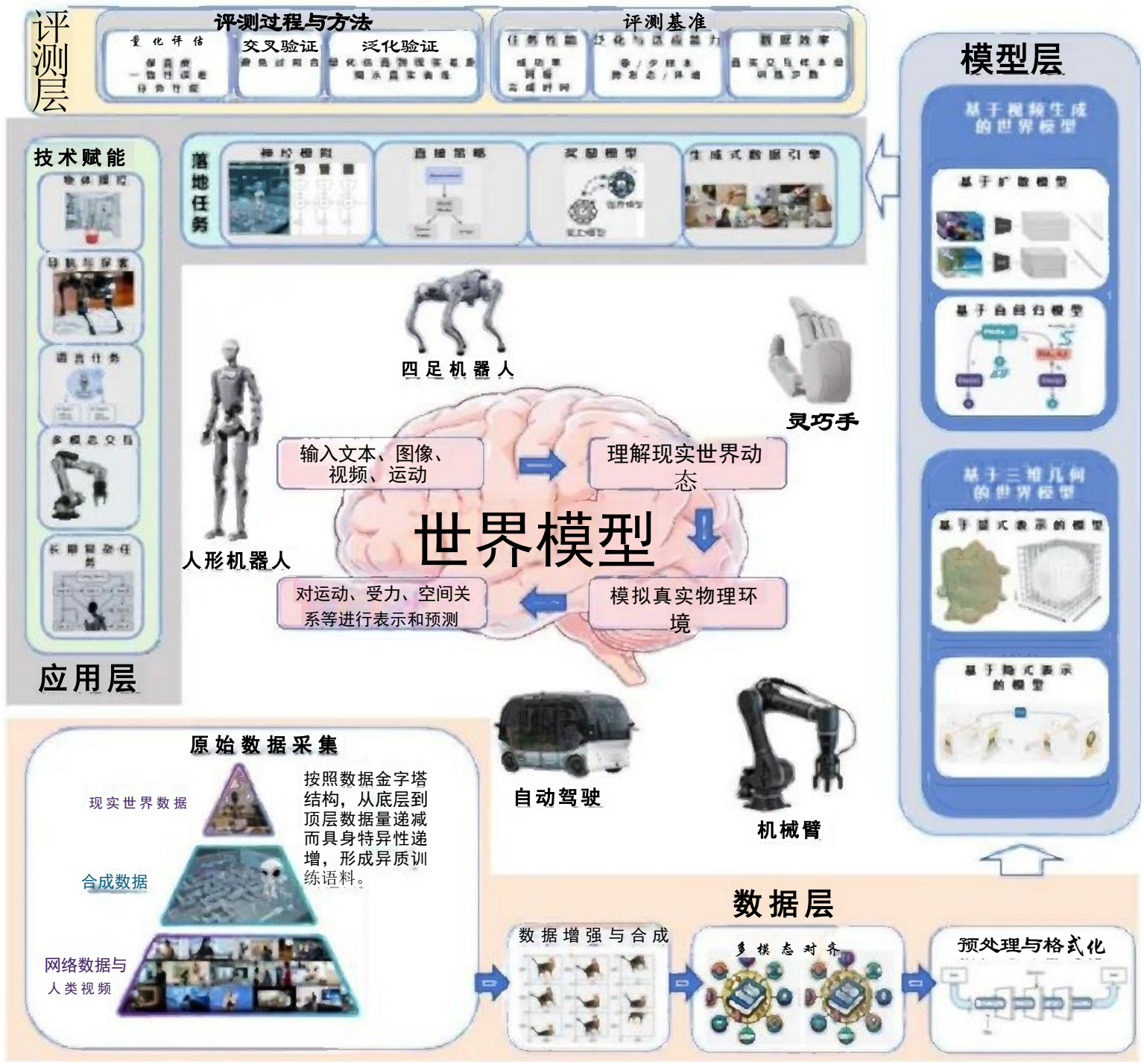

图1世界模型赋能具身智能的框架示意

准则，从而推动世界模型从技术概念走向稳健、可靠的实际应用。

（一）数据层

数据层作为具身智能世界模型构建的基础支撑，负责提供多样化、高质量的训练样本，直接决定了模型对环境动态、物理交互规律及多模态融合能力的学习效果。该层采用分层递进的数据金字塔结构，从底层到顶层，数据量逐步递减而具身特异性逐步增强，形成覆盖从广度到深度的异质训练语料体系。这一层不仅包括原始数据的采集，还通过增强、对齐和训练数据构建流水线，系统性地提升数据质量，为模型层输出经过清洗、增强、多模态对齐及格式化处理后的高质量时空序列数据，为后续模型训练和真实世界泛化奠定坚实基础。

数据采集作为数据层的第一步，采用三层递进的“数据金字塔”结构，由底层到顶层逐步从海量通用向高度具身特异性过渡，形成由广到深、由浅入深的完整梯度。“数据金字塔”的最底层汇聚海量互联网原始数据，形成基础数据底座。比如，Common Crawl[3] 提供的万亿级多模态网页内容，以及YouTube—8M、Kinetics—700等大规模视频数据集，其作用是为模型注入丰富的视觉—动作模式和多样化的人类行为表征，使模型首先掌握广泛的环境统计规律与日常动作分布。在此基础上，“数据金字塔”中层转向合

成数据。中层可以通过MuJoCo、Habitat—Sim、SAPIEN[4] 等高保真模拟器，生成带有明确物理 约束的交互序列，并作为对底层数据的补充， 扩充现实世界中难以采集的数据。这一层承接 底层的通用知识，针对真实世界中难以获取的 长尾场景、罕见事件和危险情境进行针对性补 充，从而显著扩展模型对边缘情况的覆盖和鲁 棒性。“数据金字塔”最顶层则聚焦真实具身交 互数据，直接桥接从模拟到现实的鸿沟，将底层 和中层的先验知识锚定到真实的物理世界。它 们收集来自多平台机器人的实际执行轨迹，其中 包括Bridge Dataset[5]、RoboTurk[6] 以及Open X— Embodiment[7] 等数据集。这些数据集包括RGB— D视频、本体感受和精确动作标注。这些贴合任 务本身的数据可以提升模型在实际部署时的决 策精度与迁移能力。

“数据金字塔”的三层之间层层递进、相互支撑：底层提供规模与多样性，中层注入物理真实性与长尾覆盖，顶层实现最终的具身落地与精度对齐，共同构筑起从统计先验到真实交互能力的完整数据梯度。

在数据采集的基础上，数据增强模块进一步提升数据集的多样性和模型的鲁棒性。通过空间域的随机裁剪、翻转、颜色抖动，时间域的帧插值/丢帧/序列反转以及噪声注入、视角变换、动作扰动等操作，生成大量变体样本。这些变换不仅显著扩大了有效数据量，还迫使模型学习对视角变化、光照差异、传感器噪声和动作微小扰动的不变性。

为了使不同来源的异构数据真正服务于统一的世界模型，多模态对齐模块成为关键衔接环节。它通过CLIP[8] 等多模态基础模型实现视觉、语 言和动作模态间的深度语义一致性，利用对比学 习损失将同一事件的文本描述、视频帧序列和动 作轨迹拉近正样本、推远负样本，形成共享的多模 态嵌入空间。在具身智能场景中，这种跨模态一致性显著提升了人机交互的自然度，使服务机器人能够直接响应口头指令而无需繁琐的重新编程。

预处理与格式化模块将上述所有环节串联为高效的端到端自动化流程。首先，进行严格清洗，去除模糊、分辨率低、被严重遮挡或有异常动作的样本；接着，完成精细标注与分段，生成动作边界、物体检测框和关键点；最后，通过token化或潜在表示压缩，将原始高维数据转化为适合模型输入的时空序列，保证海量异构数据的高效流入训练阶段，大幅降低人工干预成本，同时为模型提供干净、结构化且语义丰富的高质量输入。

通过“采集—增强—对齐—处理”的完整闭环，数据层不仅提供了规模庞大的训练基础，而且确保了数据的具身特异性、多样性和跨模态一致性，为世界模型在真实机器人任务中的泛化能力和物理理解能力打下了坚实的基础。

（二）模型层

模型层的设计直接决定了具身智能代理对真实世界的感知深度、理解精度与交互能力。它通过将高维、多模态感知数据高效压缩为紧凑且语义丰富的内部表征，为长期规划、不确定性建模、跨场景泛化以及物理一致性推理奠定基础。模型层为应用层提供根据不同应用场景输出的决策支持信息，如未来的视频序列预测或状态轨迹、端到端的动作序列分布或具体控制指令、密集的内部奖励标量、合成的虚拟演示轨迹或边缘场景数据。

现有具身智能世界模型主要分为两大互补范式：基于视频生成的模型和基于三维几何的模型。前者聚焦像素级时空序列预测，擅长视觉密集型动态建模；后者强调几何结构精度与物理一致性，更适用于精确交互与力学模拟场景。二者共同推动模型从纯2D视觉预测向4D结构化世界建模演进。

基于视频生成的模型通过学习图像序列的

时空动态，在像素空间直接模拟世界演化，为视觉主导的预测与规划提供支持。根据生成机制，可再次细分为基于扩散的模型和基于自回归的模型。

基于扩散的模型采用渐进去噪过程，结合因果建模、动作条件或文本—视频对齐，生成视觉逼真、时空连贯的长序列，能较好地刻画环境随机性与多模态交互。在具身智能中，它们擅长合成高质量机器人训练数据与虚拟演示轨迹，适用于离线强化学习、模拟环境构建及边缘场景扩充。尽管计算和显存开销较大，但其视觉保真度、动态连贯性与大规模数据生成能力显著提升了任务的泛化性能。

基于自回归的模型则逐帧或逐Token进行条件预测，将序列建模为严格因果链，天然适合长时序建模与逐步推理。它从海量视频中提取通用世界先验，支持多模态条件下的未来状态预测。在具身场景中，它们特别擅长高效在线规划、长期预测与物理逐步推理，实现从视觉输入到多步自主决策与零样本指令跟随的闭环。扩散模型更注重单次生成质量，自回归模型则在计算扩展性、长序列建模与因果一致性上占优，二者共同完善了像素级视频世界模型。

然而，像素级表示在精确物理交互、碰撞检测、多视角一致性及跨本体迁移方面存在局限。为此，基于三维几何的世界模型应运而生。它通过将环境显式或隐式编码为结构化的3D/4D几何表示，使世界模型在预测未来状态时不仅依赖像素外观变化，还能直接推理物体间的空间关系约束、物理连续性和多视角一致性，实现更精确、更具物理可信度的长期时空演化预测。

显式表示的模型将环境以三角网格、体素、占用网格或密集点云等结构化几何形式直接编码为世界模型的可操作输入，使得模型能够根据空间信息约束进行因果推理。EmbodiedGen[9] 1 和PointWorld[0] 作为起点，聚焦于通过多阶段生成流程创建可交互的3D资产。在此基础上，Dream2Flow' 等后续工作进一步深化，超越单纯的几何生成，转而深入探索接触密集型物理模拟与物体部件级的分解与重组。尽管受到分辨率、存储与生成复杂度的限制，但其结构化、可编辑、易集成物理引擎的特性使其在高精度交互任务中不可替代。

隐式表示的模型将环境通过连续神经场函数隐式参数化为世界模型的可微分几何表示，实现高保真新视角合成、光照建模、动态4D扩展，以及从稀疏观测到完整3D/4D世界的高质量重构。典型方法包括NeRF[12][1] 、3D Gaussian Splatting[3] 及其动态变体、GaussianWorld 等 。在具身智能中，它们特别适用于实时3D地图构建、精细物体操纵、避障导航及大规模场景重建，例如StreetSurf[15] 和GaussCtrl[16] 在城市场景中 的应用。尽管优化过程复杂，但其几何连续性、 多视角一致性、物体级可编辑性及对动态环境的 适应能力使其成为当前最具灵活性与表现力的 几何建模范式。

（三）应用层

在具身智能世界模型的整体框架中，应用层扮演着将抽象模型能力真正转化为实际具身系统解决方案的关键桥梁。它紧密衔接数据层、模型层与真实环境交互，最终使世界模型从实验室走向可落地的机器人系统。同时，应用层向评测层提供真实机器人任务的执行成功率、完成时间、碰撞次数、泛化测试结果及安全违规记录等核心运行数据。该层建立了从底层技术赋能到直观任务落地的纵向驱动逻辑：技术赋能模块直接揭示了世界模型如何系统性地辅助具身智能模型的开发与训练，而落地任务模块则通过可感知的典型场景验证了这些技术路径的有效性。

在技术赋能维度，世界模型可被用作四大研发工具。首先，可作为神经模拟器，接受当前观测和动作输入，直接生成未来视频序列或状态

轨迹。这一机制支持快速前向展开，常用于模型基规划和蒙特卡洛树搜索，从而显著提升长时序决策的效率与可行性。在此基础上，世界模型作为直接策略，代表了更激进的决策范式。该范式将世界模型直接作为策略网络，输入状态信息、文本提示或多模态观测数据，即可通过端到端推理直接输出动作序列或动作概率分布。代表性工作如Cosmos Policy[7] 及Motus18,通过世界模型潜 在空间的因果推理能力，在零样本或少样本场景 下展现出强大的泛化性能。其次，世界模型也可 作为奖励模型，利用世界模型的预测能力，将稀 疏的外部奖励转化为密集的内部奖励，支持强化 学习的高效训练，并极大地降低对人工标注的依 赖。最后，作为数据引擎，世界模型闭环数据不 断合成与增强，形成持续学习的自举机制。它一 方面可用于离线强化学习数据扩充，另一方面可 支持长尾场景覆盖与数据多样性提升。该引擎有 效缓解真实机器人交互数据稀缺问题，为预训练 与微调持续注入高质量合成样本。

通过神经模拟器→直接策略→奖励模型→数据引擎的层层递进，技术赋能模块构建起从“内心预演”到“即时行动”、从“奖励自监督”到“数据自举”的完整工具链，为落地任务提供了坚实的技术底座。

应用层的落地任务则是这些技术赋能的最终检验与体现，涵盖从基础到高级的多层次、多场景应用需求，具体聚焦物体抓取、导航探索、语言条件任务、多模态交互以及长期复杂任务五大类。物体抓取任务首先落地，利用世界模型预测抓取后的物理交互，支持从桌面整理到工业零件装配等场景。导航探索任务则进一步扩展到移动机器人，聚焦实现动态避障与未知区域的自主巡航。在此基础上，语言条件任务引入自然语言驱动，机器人结合视觉理解与语言解析生成对应动作序列，依赖世界模型的多模态推理能力，实现零样本或少样本指令跟随。多模态交互任务再向上跃升，整合视觉、语言、触觉甚至力反馈，支持更自然的协作式物体传递或精细装配。最终，长期复杂任务代表最高挑战，涉及多步序列决策与长期规划。世界模型通过模拟长时序动态演化与不确定性建模，确保任务持续性和鲁棒性，避免累积误差导致失败。

这些任务由单步到多步、由单模态到多模态、由短期到长期，形成完整的应用梯度。通过真实机器人平台的部署验证，世界模型逐步缩小从模拟到实物的迁移差距，展现出强大的跨场景泛化能力，为具身智能走向社会化应用奠定坚实基础。

（四）评测层

在具身智能领域，建立全面的评测框架至关重要。评测层的设计涵盖评测方法、过程及基准数据集，旨在系统量化世界模型的生成质量、物理一致性以及在下游具身任务中的实际适用性，从而为模型迭代和实用落地提供可靠依据。在与其余各层的交互中，评测层指出数据分布的偏差，量化生成内容的物理违背情况或指令跟随偏差，指导模型架构优化或超参数调整，并且提供识别系统在极端条件下的失效模式，设定安全阈值，推动应用策略的保守化调整或安全护栏的部署。

评测过程采用分阶段的系统化流水线架构，整体划分为量化评估阶段、泛化验证阶段。首先进入量化评估阶段，通过计算视觉保真度、物理一致性误差以及下游任务性能等多维度指标刻画模型的核心能力。这一阶段强调交叉验证和A/B测试设计，以有效避免过拟合偏差。在此基础上，评测自然过渡到泛化验证环节：在分布外数据或真实机器人平台上重新运行相同评估流程，从而量化仿真—现实差距并揭示模型在真实世界中的真实表现。通过这一从内在指标到外在应用的逐步桥接，评测层真正确保了模型在复杂具身任务中的实用价值。

基于上述过程，评测基准按性能维度进行分类，可分为生成质量基准、任务表现基准、泛化能

力基准、数据效率基准以及安全性基准五个核心方面。其中，视觉保真度主要通过弗雷歇Inception距离（FID）和弗雷歇视频距离（FVD）[19]来量化生成内容与真实分布的相似程度，它们基于深度特征统计分布的距离度量，能够超越传统的像素级误差，更敏锐地捕捉生成图像或视频在纹理细节、色彩自然度以及时序连贯性上与真实数据分布的深层相似性，从而客观反映“看起来是否逼真”。其次，为了使评估更具针对性和可比性，行业常选用不同侧重的代表性数据集进行测试，比如UCF—101作为经典的侧重动作识别视频数据集，更侧重考察模型对人类动作的流畅性和主体一致性的识别能力。在此基础上，**Something-Something-v2**[20]则进一步提升难度，聚焦物体间的细粒度交互，能更真实地反映模型对物体运动规律、时空连续性和交互细节的捕捉能力。物理一致性评估进一步深入，关注物体持久性、因果准确率等关键属性：验证运动轨迹的合理性，Physion 基准[21]通过生成轨迹与真实物理模拟的偏差检验碰撞、滚动、重力等行为。为了应对其他不合理现象，GAIA-122 引入物理违反率量化物体穿模、异常加速度等的出现情况。当生成质量得到验证后，评测的重点便自然转向任务性能基准，考察世界模型作为规划器在下游具身任务中的实际效果。再次，泛化与适应能力基准可以用来检验模型的迁移潜力，包括零/少样本成功率、跨形态迁移和仿真—现实差距。其中，Libero[23] 作为起点，提供跨任务组合评估。在此基础上，RealWorld RL Suite和Open X-Embodiment支持跨数据集、跨机器人的迁移验证。RT-224 与从零训练基线的对比，将评估推向最高难度：跨机器人形态的零样本或少样本迁移。在此之上，数据效率基准则关注模型在资源受限场景下的实用性，比如，可以选择在DMControl[25] 和MetaWorld[126]上对比所需真实交互样本量与训练步数。最后，鲁棒性与安全性基准考察模型在极端条件下的稳健表现，包括对抗扰动鲁棒性、失败恢复率以及伦理合规性。通过从生成质量到任务表现、泛化能力、数据效率，再到鲁棒性与安全性的层层评价，构建起一个完整、互补的评价体系。

三、世界模型赋能具身智能的行业落地

世界模型驱动的具身智能正成为破解具身智能行业落地瓶颈的核心引擎。以“高保真物理模拟+长时序因果预测”为技术基座，世界模型通过深度融合多模态轨迹、海量视频先验与物理规律，推动具身智能从“反应式执行”转向“预测式规划”、从“任务特定适应”升级为“开放场景泛化”。本文围绕工业制造中的柔性装配与智能质检、物流仓储中的非结构化分拣与路径优化等核心维度，系统阐述世界模型如何赋能具身智能企业在“数据—模型—场景—反馈”闭环中构建可靠的产业护城河，实现从“实验室演示”到“万台规模部署”、从“结构化任务”到“开放世界自主”的跨越式升级。

（一）世界模型赋能具身智能在工业制造行业落地

传统工业制造面临多品种小批量生产、非结构化物料堆放、工艺公差波动以及人力密集型装配等核心瓶颈127。在装配阶段，通常需协调毫米级定位精度、力控柔顺性与动态环境适应等多维度约束，导致部署周期长、缺陷易流入下游环节、柔性切换成本高昂。世界模型通过整合多模态轨迹数据、物理先验及历史交互序列，结合高保真数字孪生与预测式规划技术，可构建内部模拟环境，提前预判接触力学状态、物体变形与碰撞风险。同时，基于因果推理与长时序优化算法自动生成符合物理约束的运动轨迹，将传统数周的实训迭代周期压缩至小时级，显著降低后期返工率与安全隐患28】。

星动纪元的星动L7落地[2] 29印证了世界模型在解决工业制造核心难题中的价值。汽车与3C电子制造长期受制于工件形变、公差累积与无序抓取的物理复杂性，例如部件变形、装配应力分布等参数难以精准建模。星动L7搭载端到端原生机器人大模型ERA-42，通过海量互联网视频与自采集交互数据联合预训练，在指令微调阶段构建物理动态预测与长时序规划能力。其核心世界模型模块通过高效的潜在空间表征与实时推理压缩，可在产线实时融合“视觉—力觉—本体”感受数据，预测装配接触力学状态并在线生成柔顺阻抗轨迹，指导机器人完成螺钉钻入、部件插入、精密组装等精细操作。如图2所示，虽然星动L7是一个工程化全尺寸双足人形机器人，但在效率与鲁棒性、清晰性与直接性方面与前沿模型相当，并在实用性与即时可用性、逻辑流程一致性、专家级沟通、使用实例特异性等关键指标上明显领先。截至2026年初，星动L7已在制造场景实现试点与小批量部署，任务完成效率得到大幅度提升，形成“场景越丰富，模型越智能”的正向循环，标志着世界模型从实验室预测工具向产业级决策引擎的实质跨越。

（二）世界模型赋能具身智能在物流行业落地

当前，物流仓储场景正面临从自动化向智能化升级的关键瓶颈。传统自动化系统虽能执行预设任务，但普遍存在环境适应能力差、任务切换效率低、异常处理依赖人工干预等问题。尤其在面对海量SKU、动态订单波动及非标操作时，传统机器人仅能依赖预设规则与有限感知进行工作，导致分拣准确率在复杂场景下显著下降，柔性部署成本高昂。以世界模型为核心的具身智能技术，正通过构建高保真的物理场景模拟与预测能力，为物流机器人赋予“想象未来”与“因果推理”的关键认知能力。

具体而言，系统通过融合激光雷达、深度相机等多模态传感器数据，实时构建并更新一个可预测的数字化仓储环境。这个世界模型不仅能精准反映当前货架位置、货物状态、机器人及人员分布等静态与动态要素，而且能以前瞻视角，高保真地推演未来数秒至数十秒内各元素的交互与状态演变，例如预测特定区域即将形成的交通拥堵、评估不同任务分配策略下的全局作业效率，甚至模拟突发设备故障可能引发的连锁反应。

以业内领先的极智嘉（Geek+）智能仓储分拣场景的实际应用[3 2]为例，创新性引入世界模型技术，以应对海量SKU、高波动订单下的高效精准分拣挑战，印证了该技术在破解物流“动态复杂度”方面的独特价值。针对传统方案中，移动机器人与拣选站协同效率受限于固定路径、静态

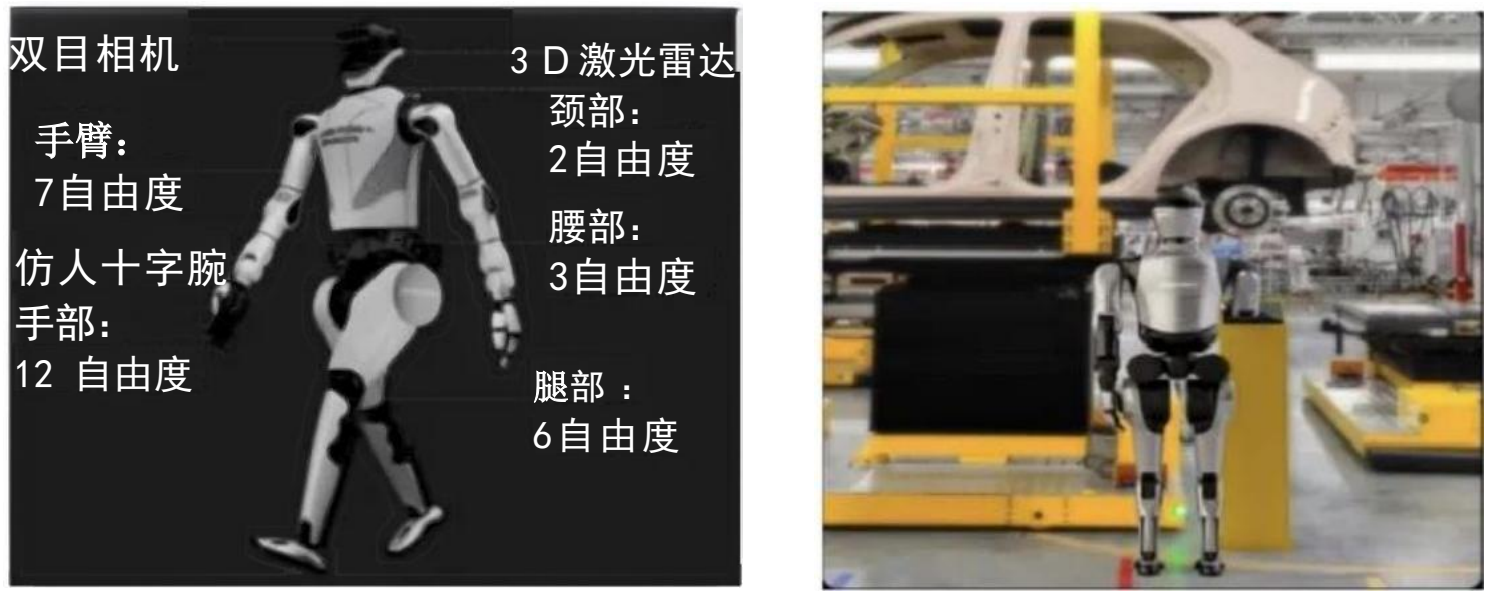

图2星动纪元的星动L7 搭载ERA-42 大模型在完成汽车密组装任务130]

任务分配，以及面对新SKU或临时仓位变更时适应迟缓的痛点，极智嘉构建了“具身智能体—仓储世界模型”协同系统。该系统通过三维视觉与激光雷达实时感知仓内全局状态，并将其输入至预训练的视频预测世界模型中。该模型在大量历史运行视频与仿真数据上训练，能够高保真地预测未来数秒至数十秒内仓内关键元素的动态演变，如预测某拣选工作站即将出现拥堵、空载自动导引车（AGV）到达目标货架的最优无冲突路径以及人工拣选员完成当前操作的大致时间。如图3所示，中央调度系统不再是简单的任务派发器，而是基于世界模型提供的多步未来场景“推演”，进行前瞻性优化决策。

四、世界模型赋能具身智能的机遇与发展趋势

随着具身智能向真实世界的规模化落地迈进，世界模型已成为连接感知、推理与行动的关键认知引擎。然而，其深度应用仍面临物理失真、指令偏差、安全风险不可控及评估体系缺失等挑战，制约从仿真到现实的可靠迁移，亟须在物理真实性、多模态对齐、安全约束与任务评估四个关键方向实现突破，构建安全、可信、高效且高度任务对齐的技术体系。

（一）物理可信生成，筑牢具身智能安全基石

当前，视频生成模型在具身智能与机器人应用中面临一大核心问题：生成内容违背物理规律，如违反牛顿运动定律、动量守恒、质量守恒等，具体包括倒水时杯中液面不变，或物体相互穿透。为了应对上述挑战，可以尝试采用多种解决路径。一是引入物理先验，如基于哈密顿或拉格朗日力学构建动力学约束，或结合刚体仿真引擎生成符合物理的粗略轨迹，再由扩散模型优化视觉质量。二是通过视觉语言模型（VLM）/ 大语言模型（LLM）解析场景中的物体属性与交互语义，生成更详尽的物理提示以引导视频生成。三是融合可供性。从人类操作视频中学习“可交互区域”，并将接触热点演化预测作为生成条件，以提升动作合理性。然而，现有方法多为拼接式方案，依赖外部模块，易引入伪影且通用性差。未来研究应聚焦于使视频模型原生内化物理定律，通过新型架构设计与训练范式，从根本上提升生成内容的可信度与安全性。

（二）精准指令跟随，夯实具身智能训练数据根基

当前，文本到视频生成模型在执行用户指令方面仍存在显著短板：尽管能识别提示中的主体对象，却常难以准确提取并执行指定动作。此外，模型难以按需生成静态镜头，且难以在视频

图3极智嘉移动机器人的智能仓库分拣任务应用[31]

中可靠地嵌入文字标注。为了提升指令的遵循能力，近期研究尝试引入多模态条件引导：一是条件引导的世界模型，通过将自然语言指令或动作序列融入潜空间，实现更忠实的任务执行以及精细动作序列控制。二是偏好对齐微调。借鉴基于人类反馈的强化学习（RLHF）范式，利用人类偏好数据集对模型进行指令级细调，增强语义忠实度。然而，这些方法多依赖外部组件进行指令语义对齐。未来方向应聚焦于原生语义内化，通过端到端架构设计与自监督对齐训练范式，赋予模型内在的任务理解能力，从而为具身智能提供语义精确、结构可靠的合成数据。

（三）构建机器人导向的评估体系，推动世界模型实现高质量发展

当前，视频生成模型在具身智能中的应用面临评估标准缺失的瓶颈：主流指标多聚焦于感知质量或语义对齐度，难以反映其在机器人任务中至关重要的物理一致性、动作准确性与预测可靠性。由于缺乏统一、量化的评测框架，研究者常依赖下游任务代理指标或主观人工评分，前者间接且场景受限，后者则成本高、可复现性差。尽管已有尝试，如WorldModelBench[33]引入视觉语言模型（VLM）裁判评估指令遵循与常识合理性，EWMBench[34]提出场景一致性、运动正确性与语义质量三维框架；VideoPhy[35] 物理常识基准则聚 焦动力学违反检测。但这些方法仍难以覆盖精细 操作场景所需的细粒度评判能力，且多数受限于 仿真环境的视觉与物理保真度。未来亟需建立以 机器人任务为中心的多维评估体系——融合3D场 景重建、动力学验证与任务成功率映射，开发可 量化、可扩展、任务相关的自动化评测管道。

（四）筑牢安全生成防线，护航具身智能可信交互

当前，视频生成模型在安全机制方面严重滞后，缺乏有效的防护护栏，易生成包含暴力、违法、隐私泄露或误导性内容的视频。尽管大语言模型已建立了较成熟的对齐与拒答机制，但面向视频模态的安全技术仍处于早期探索阶段——现有方案如近期研究尝试构建专用安全机制。一是内容过滤与红队测试：如ASIMOv[6] 安全基准 提供大规模数据集与连续评估协议，聚焦物理安 全与语义安全的量化检测。二是前瞻式预演：利 用世界模型自身生成潜在未来轨迹，并在潜空间 或视频输出中部署分类器，以检测冲突，实现主 动拦截。三是偏好对齐扩展：借鉴基于人工智能 反馈的强化学习（RLAIF)[37], 将人类/代理安全 偏好融入训练，抑制分布内的有害行为。然而， 这些方法大多通用性不足、在长尾分布外场景中 的鲁棒性有限。未来研究亟须构建统一的安全生 成框架，开发可扩展的多维度护栏，并建立以机 器人交互为导向的综合安全基准，实现从“可生 成”向“安全可控生成”的范式转变，为具身智 能在开放环境中的可信部署提供坚实保障。

五、总 结

综上所述，本文系统阐述了世界模型作为具身智能核心引擎的关键作用。通过构建环境模拟与预测推演能力，世界模型有效破解了数据稀缺、泛化不足与实时决策等瓶颈，推动具身智能从受限场景迈向开放环境的自主行为生成。围绕数据、模型、应用与评测四层体系，本文结合工业制造、物流等场景实践，揭示了其从实验室走向规模化部署的路径与价值。世界模型与具身载体的深度融合将持续拓展智能系统的认知边界，为构建通用行动智能奠定基础，助力我国在智能化浪潮中形成技术引领与产业优势。

然而，本研究仍存在一定的局限性。本文提出的“数据—模型—应用—评测”框架尽管在逻辑上形成了闭环，但在实际工程落地中，各层级间的接口标准尚未统一，跨平台迁移的兼容性与实时推理的算力成本问题在文中仅作定性探讨，

缺乏具体的能效比数据支撑。展望未来，后续研究将重点聚焦于以下三个方向：一是构建开源共享的具身世界模型基准平台，联合产学研各方力量，建立包含多模态物理交互、长时序因果推理及极端安全场景的标准数据集与评测沙箱，降低行业研发门槛并加速技术迭代；二是探索轻量化与边缘侧部署技术，针对机器人本体算力受限痛点，研发模型蒸馏、动态稀疏化及端云协同推理架构，实现世界模型在低功耗嵌入式设备上的毫秒级实时响应；三是深化安全对齐与伦理治理研究，从算法底层植入物理约束与价值对齐机制，开发可解释性强的决策追溯系统，确保具身智能在开放人机共融环境中的行为可控、可信且符合伦理规范，最终推动具身智能产业从“单点突破”迈向“生态繁荣”。

参考文献

[1]共产党员网. 党的二十届四中全会《建议》学习辅导百问：为什么要全面实施“人工智能+”行动?[EB/OL].(2025-12-05)[2026-02-04].https://www.12371.cn/2025/12/04/ART11764857025723809.shtml.

[2]新华社. 超150家企业!具身智能“跑”入更多生活场景[EB/OL].(2025-12-27)[2026-02-05].https://www.news.cn/20251227153a4116e5f6046659ff1b3dba037285e/c.html.

[3]PATELJ M.Introduction to common crawl datasets[M]// Getting structured data from the internet running web crawlers/scrapers on a big data production scale.Berkeley,CA:Apress,2020:277-324.

[4]XIANG F,XU ZY,LI D,et al.SAPIEN:A simulated part-based interactive environment[C]V/Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle:IEEE2020.

[5]EBERT F,LIU K,FINN C,et al.Bridge data·Boosting generalization of obotic skills with eross-domain datasets[EB/OL].(2021-09-13)[2026-04-27].https://arxiv.org/abs/2109.13396.

[6]MANDLEKAR A,ZHU Y,GARG A,et al.RoboTurk:A crowdsourcing platform for robotic skill learning through imitation[CV/Conference on Robot Learning.PMLR,2018.

[7]O'NEILL A,REHMAN S,MADDUKURI A,et al.Open X-Embodiment:Robotic learning datasets and RT-X models:Open X-Embodiment collaboration 0[C]//2024 IEE International Conference on Robotics and Automation. Yokohama:IEEE.2024.

[8]RADFORD A,KIMJW,HALLACY C,et al.Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. [S.1.]:PMLR,2021.

[9]WANG X,LIU L,CAO Y,et al.EmbodiedGen:Towards a generative 3D world engine for embodied intelligence[EB/OL].(202-06-10)[2026-04-27].https//arxiv.org/abs/2506.10600.

[10]HUANG W,CHAO YW,MOUSAVIAN A,et al.PointWorld:Sealing 3D world models for in-the-wild robotic manipulation [EB/OL](2026-01-03[2026-04-27]https//arxivorg/als/260103782

[11]DHARMARAJAN K,HUANG W,WU J,et al. Dream2Flow:Bridging video generation and open-world manipulation with 3D object flow[EB/OL].(2025-12-24)[2026-04-27].https://arxiv.org/abs/2512.24766.

[12]MILDENHALL B,SRINIVASAN P P,TANCIK M,et al.NeRF:Representing scenes as neural radiance fields for view synthesis[J].Communications of the ACM,2021,65(1):99-106.

[13]KERBL B,KOPANAS G,LEIMKUHLER T,et al.3D Gaussian splatting for real-time radiance field rendering[J].ACM Transactions on Graphics,2023,42(4):139.

[14]ZUO S,ZHENG W,HUANG Y,et al.GaussianWorld:Gaussian world model for streaming 3D occupancy prediction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.Nashville:IEEE,2025.

[15]CUI X,YE W,WANGY,et al.StreetSurfGS:Scalable urban street surface reconstruction with planar-based Gaussian splatting[J].IEEE Transactions on Circuits and Systems for Video Technology,35(9):8780-8793.

[16]WU J,BIAN JW,LI X,et al.GaussCtrl:Multi-view consistent text-driven 3D Gaussian splatting editing[C]//European Conference on Computer Vision.Cham:Springer Nature Switzerland,2024.

[17]KIM M J,GAO Y,LIN T Y**,et al.Cosmos Policy:Fine-tuning video models for visuomotor control and planning[EB/OL].**(2026-01-16)[2026-04-27].https://arxiv.org/abs/2601.16163.

[18]BI H,TAN H,XIE S**,et al.Motus:A unified latent action world** model[EB/OL].(2025-12-13)[2026-04-27].https://arxiv.org/abs/2512.13030.

[19]SOOMRO K.ZAMIR A R,SHAH M.UCF101:A dataset of 101

human actions classes from videos in the wild[EB/OL].(2012-12-04)[2026-04-27].https:/larxiv.org/abs/1212.0402.

[20]GOYAL R,EBRAHIMI K S,KAHOU SEet al.The "something something"video database for learning and evaluating visual common sense[C]//Proceedings of the IEEE International Conference on Computer Vision.Venice:IEEE,2017.

[21]BEAR D M,WANG E,MROWCA D,et al.Physion Evaluating physical prediction from vision in humans and machines[EB/OL].(2021-06-08)[2026-04-27].https://arxivorg/abs/2106.08261.

[22]HU A,RUSSELL,YEO H,et al.GAIA-1:A generative world model for autonomous driving[EB/OL].(2023-09-17)[2026-04-27].https://arxiv.org/abs/2309.17080.

[23]LIU BZHU Y,GAO C,et al.LIBERO:Benchmarking knowledge transfer for lifelong robot learning[CV/Advances in Neural Information Processing Systems.New Orleans:Curran Associates,2023:44776-44791.

[24]BANSAI H,LIN Z,XIE T,et al.RT-2:Vision-language-action models transfer web knowledge to robotic control[EB/OL].(2023-07-15)[2026-04-27].https://arxiv.org/abs/2307.15818.

[25]HAFNNER D,LILLICRAP T,NOROUZI M,et al.Mastering Atari with discrete world models[EB/OL].(2020-02-19)2026-04-27].https://arxiv.org/abs/2010.02193.

[26]YU T,QUILLEN D,HE Z,et al.Meta-World:A benchmark and evaluation for multi-task and meta reinforcement learning[CVI Conference on Robot Learning.[S1.]:PMLR,2020.

[27]胡亚男，焦艳红，田思苗，等. DeepSeek驱动的中小企业数智化转型：低成本AI技术赋能的转型路径[J].数字化转型，2025,2(8):5-12.

[28]光明日报. 以高水平科技自立自强引领发展新质生产力[EB/OL].(2025-11-25)[2026-02-04].https://www.news.cn/politics/20251125/c4cf542eb4cb461d85edcf8abddfacOa/c.html.

[29]北京星动纪元科技有限公司. 星动L7全尺寸双足人形机器人(新一代)[EB/OL].(2025-12-11)[2026-04-27].https://www.robotera.com/robot/17.html.

[30]北京星动纪元科技有限公司. 应用场景服务千行百业万户[EB/OL](2025-11-01)[2026-04-27].https:l/www.robotera.com/application.html.

[31]极智嘉(Geek+).极智嘉发布全新通用机械臂操作技术方案，破解仓储超大规模商品拣选难题[EB/OL].(2025-08-27)[2026-04-27].https://www.geekpark.net/news/353138.

[32]极智嘉(Geek+).智能分拣解决方案[EB/OL].(2025-10-18)[2026-04-27].https://www.geekplus.com/zh-cn/solutions/sorting.

[33]LI D,FANG Y,CHEN Y,et al.WorldModelBench:Judging video generation models as world models[EB/OL].(2025)[2026-04-27].https://*arxiv*.org/abs/2502.20694.

[34]YUE H,HUANG S,LIAO Y,et al.EWMBench:Evaluating scene,motion,and semantic quality in embodied world models[EB/OL](2025-05-09)[2026-04-27].https//arxiv.org/abs/2505.09694.

[35]BANSAL H,LIN Z,XIE T,et al.VideoPhyEvaluating physical commonsense for video generation[EB/OL].(2024)[2026-04-27].https://arxiv.org/abs/2406.03520.

[36]CHEN Z,PINTO F,PAN M,et al.SafeWatch:An efficient safety-policy following video guardrail model with transparent explanations[EB/OL].(2024-12-06)[2026-04-27].https://arxiv.org/abs/2412.06878.

[37]LIU J,LIU G,LIANG J,et al.Improving video generation with human feedback[EB/OL].(2025-01-13)[2026-04-27].https://arxiv.org/*abs*/2501.13918.

作者简介：

宫丽娜，南京航空航天大学，副教授，博士，研究方向：大模型原理与技术；

徐嘉龙，南京航空航天大学，硕士，研究方向：具身智能；

朱哲，南京航空航天大学，博士，研究方向：三维视觉；

李安，南京航空航天大学，硕士，研究方向：三维视觉；

赵彦超(通信作者)，南京航空航天大学，博士，研究方向：大数据处理，边缘计算，计算机网络，电子邮箱：yczhao@nuaa.edu.cn。