中图法分类号:TP18;TP24　　**文献标识码**:A　　**文章编号**:1006-8961(2026)08-2705-26

论文引用格式:Li C Y, Zhang J B, Xiao J H, Yan B W, Guo S Y, Ye T R, Lin W S and Zhai G T. 2026. Embodied intelligence model evaluation: from perception to execution. Journal of Image and Graphics, 31(8):2705-2730(李春一, 张建博, 肖嘉豪, 闫博闻, 郭晟毓, 叶桐瑞, 林维斯, 翟广涛. 2026. 具身智能模型评测:从感知到执行. 中国图象图形学报, 31(8):2705-2730)[DOI:10.11834/jig.250550]

具身智能模型评测:从感知到执行

李春一[1,2],张建博[1,2],肖嘉豪[1,2],闫博闻[1,3],郭晟毓[1,4],叶桐瑞[1,5],
林维斯[6],翟广涛[1,2*]

1. 上海人工智能实验室评测专项组,上海　200032; 2. 上海交通大学信息与电子工程学院,上海　200240;
3. 清华大学深圳国际研究生院数据与信息研究院,深圳　518055; 4. 北京大学元培学院,北京　100871;
5. 同济大学应用心理学系,上海　200292; 6. 南洋理工大学计算机与数据科学学院,新加坡　639798,新加坡

摘　要:具身智能通过"身体—环境—任务"闭环交互,被视为迈向通用人工智能的核心路径。然而,与模型参数和训练数据的高速扩张形成鲜明对照,其评测体系仍处于"任务碎片化、指标多元化、平台封闭化"的自发状态,造成同一功能在不同研究中的性能差异超过10%,却无法判定来源是算法创新、数据扩充还是评测偏差,严重阻碍了技术迭代与产业落地。本文围绕"从感知到执行"的完整链路,对2020—2025年间发表于CVPR(IEEE/CVF Conference on Computer Vision and Pattern Recognition)、ICRA(IEEE International Conference on Robotics and Automation)、NeurIPS(Conference on Neural Information Processing Systems)等顶级会议与Nature、IJRR(International Journal of Robotic Research)、JMLR(Journal of Machine Learning Research)等期刊的百余篇文献进行系统梳理,首次提出"静态数据集—仿真平台—真实机器人"三级金字塔式评测范式,并从感知、认知、决策、执行4个环节拆解出20余项核心能力维度与量化指标,分别从方法学假设、适用边界、固有局限和成本—可信度曲线4个角度进行横向对比。本文进一步汇总了46套主流基在数据规模、任务类型、评估指标、开源程度和安全伦理考量等维度的差异。基于此,提出"能力导向、协议统一、三级协同"的未来框架:1)在任务层,由"功能对标"转向"能力对标",建立可分解、可溯源、可加权的多维误差体系;2)在协议层,制定统一的场景描述、接口规范与指标定义,实现跨平台、跨任务、跨模型的可比性;3)在系统层,构建可远程接入、7×24 h运行的共享机器人集群,形成"仿真预训练—真机微调—在线更新"的闭环追踪,降低重复建设成本,打通从实验室到场景落地的"最后一公里"。最后讨论了安全伦理、文化偏见和能效评估等新维度如何纳入量化框架,并给出标准化路线图的短、中、长期目标。本文工作为具身智能从"技术涌现"走向"科学共识"提供了可操作的评测基础设施与参考范式,具体的静态—仿真—真机榜单展示在:https://opencompass.org.cn/embodied-intelligence。

关键词:具身智能;评测体系;感知—认知—决策—执行;仿真平台;真机测试;仿真—现实一致性

Embodied intelligence model evaluation: from perception to execution

Li Chunyi[1,2], Zhang Jianbo[1,2], Xiao Jiahao[1,2], Yan Bowen[1,3], Guo Shengyu[1,4], Ye Tongrui[1,5],
Lin Weisi[6], Zhai Guangtao[1,2*]

1. Center of AI Evaluation, Shanghai AI Lab, Shanghai　200032, China; 2. School of Information Science and Electronic Engineering, Shanghai Jiao Tong University, Shanghai　200240, China; 3. Institute of Data and Information, Tsinghua Shenzhen International

收稿日期:2025-11-05;**修回日期**:2026-01-29;**预印本日期**:2026-02-05

*通信作者:翟广涛　zhaiguangtao@pjlab.org.cn

基金项目:国家自然科学基金项目(62225112,625B2118);新加坡教育部基金项目(MOE-T2EP20123-0006)

Supported by:National Natural Science Foundation of China(62225112,625B2118);Singapore MoE Foundation(MOE-T2EP20123-0006)

Graduate School, Shenzhen 518055, China; 4. Yuanpei College, Peking University, Beijing 100871, China; 5. Department of Psychology, Tongji University, Shanghai 200292, China; 6. College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore

Abstract: Embodied intelligence refers to agents that sense, reason, act, and learn through a tight "body-environment-task" feedback loop. It is widely regarded as the most promising route to general-purpose artificial intelligence (AI). Paradoxically, although model parameters, training data, and compute budgets have grown by four orders of magnitude in only five years, the evaluation ecosystem remains fragmented and largely irreproducible. The same functional capability (e. g. , "pick-and-place") is formalized as object-detection mean average precision in one paper, as 3D intersection over union in a second, as task success rate in a third, and as human preference score in a fourth, with absolute gaps that exceed 10%—15% among "state-of-the-art" results. Environments, random seeds, physics parameters, sensor noise models, and success criteria are rarely open-sourced, and thus, the community cannot tell whether an apparent improvement originates from algorithmic innovation, data scale, evaluation cherry picking, or simple benchmark over-fitting. This uncertainty has become a critical bottleneck for scientific progress and industrial adoption. To address the gap, we conduct a systematic survey of 100+ papers published in IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE International Conference on Robotics and Automation, Conference on Neural Information Processing Systems, Robotics: Science and Systems, International Conference on Learning Representations, International Journal of Robotic Research, and Journal of Machine Learning Research between 2020 and 2025. Then, we present the first holistic analysis of embodied AI evaluation. We organize the landscape into a three-tier pyramid: static datasets, simulation benchmarks, and real-robot trials. Each tier trades off cost, controllability, and ecological validity. Across all tiers, we decompose evaluation space into four tightly coupled stages: perception, cognition, decision, and execution. Then, we identify fine-grained capability dimensions along with 20+ quantitative metrics. For every dimension, we analyze the following: 1) the underlying methodological assumptions, 2) the boundary conditions under which the metric is valid, 3) the intrinsic limitations (e. g. , sim-to-real physics mismatch, dataset bias, human-label variance), and 4) the cost-credibility curve that determines when a researcher should ascend from static to simulation, and finally to the real world. All evaluations in this study are based on a theoretical "perception-cognition-decision-execution loop" framework. In embodied AI, perception is the process by which an agent converts raw, asynchronous, and often noisy signals from heterogeneous physical sensors, such as RGB-D cameras, LiDAR, tactile skins, joint encoders, force/torque, and inertial measurement unit units, into a registered, time-aligned digital twin that supplies metric 3D geometry and semantic labels. Cognition is then built on this torrent to distill structured, memory-enabled, and generalizable world models that encode object affordances, physical common sense, and relational and causal links, grounding free-form human language into the resulting latent space. Decision-making combines these world models with an explicit task goal and searches across symbolic, kinematic, and dynamic levels to output a coherent, risk-aware, and resource-efficient plan that is expressed as a hierarchy of subgoals, trajectories, and low-level set points. Finally, execution transforms this plan into real-world mechanical energy through high-rate feedback controllers that track desired motions or forces while compensating for actuator nonlinearities, shock, thermal limits, and human safety constraints, closing the perception-cognition-decision-execution loop that defines embodied AI. In contrast with humans, whose perception-cognition-decision-execution loop is seamlessly integrated, currently available embodied agents suffer structural flaws in every stage. Consequently, evaluation must decouple the four blocks and score accuracy, error, and latency separately to pinpoint bottlenecks. Static tests, i. e. , "written exams", run on a GPU with a few megabytes of images or point clouds in minutes, but erase dynamics and physics. Simulation, i. e. , the "job interview", adds interactive physics in tens of seconds but still drifts away from reality. Real robot trials, i. e. , "practical tests", cost thousands, take days to deploy and minutes per run, but expose every sensing drift, prediction error, or control delay, offering the ultimate ground truth before field deployment. Therefore, by integrating the three-tier pyramid of static test-simulation-real robot evaluation into the closed perception-cognition-decision-execution loop, this work delivers the first systematic quantification of coverage, cost, and accuracy trade-offs across more than 100 embodied AI benchmarks, establishing a reproducible, traceable, and extensible evaluation baseline that not only exposes the exact stage at which state-of-the-art systems

fail, but also provides a principled road map for allocating research budgets, standardizing protocols, and certifying safety before real-world deployment, marking a pivotal transition from empirical competition to scientific measurement, and accelerating the field's development toward trustworthy, deployable, and general-purpose embodied intelligence. The detailed static-dynamic-real embodied intelligence evaluation can be accessed at https://opencompass.org.cn/embodied-intelligence.

Key words: embodied intelligence; benchmark evaluation; perception-cognition-decision-execution; simulation; real-world validation; sim2real consistency

0 引　言

具身智能(embodied intelligence)作为人工智能(artificial intelligence, AI)领域最具挑战性的前沿方向之一,其核心思想在于强调智能并非仅仅源于抽象的计算或符号处理,而是深深植根于物理实体与动态环境之间的持续交互之中(Pfeifer 和 Bongard, 2006)。这一范式认为,一个智能体的认知、学习和决策能力,从根本上受到其身体形态、感知运动系统以及与环境互动方式的塑造(Anderson, 2003)。具身智能体(embodied agent)通常定义为一个能够通过传感器感知环境,并通过执行器对环境施加影响的物理实体,其智能行为在感知与决策的连续循环中,逐步达到涌现和演化(Duan 等, 2022)。

具身智能与传统人工智能在智能获取方式上存在本质区别。传统人工智能,尤其是在大数据和深度学习驱动下取得巨大成功的领域,如计算机视觉(computer vision, CV)和自然语言处理(natural language processing, NLP),其模型训练主要依赖于大规模、预先收集并标注的静态数据集(LeCun 等, 2015)。这种范式使得模型能够在特定、受限的环境中表现出色,但一旦面临开放、动态和非结构化的真实世界,其泛化能力和鲁棒性便会受到严峻挑战(Torralba 和 Efros, 2011)。相比之下,具身智能体通过与环境的实时互动来获取训练数据、检验假设并优化行为策略(Kaelbling 等, 1996)。这种学习方式使得智能体能够持续适应环境变化,并从交互中自主发现和学习新的概念与技能。

随着具身智能在研究和实际应用中的广泛使用,对其进行有效评测变得愈发重要。然而,与模型规模、参数更新频率及场景覆盖度的高速扩张形成鲜明对照的是,评测体系仍未统一,主要表现为:

1)任务定义碎片化。同一"抓取—放置"功能在不同工作中被形式化为对象检测精度、动作成功率、路径效率或人机交互满意度等异质指标,缺乏统一的操作化描述。

2)环境配置封闭化。近八成已发表具身智能模型依赖私有仿真器、非公开真机平台或定制传感器套件,随机种子、物理参数、渲染管线与噪声模型均未开源,导致结果无法复现。

3)评价协议差异化。静态数据集、高保真仿真与真实机器人,即静态、仿真和真机 3 种评测范式在成本与保真度上,呈现明显的两难;研究者往往基于可承受资源而非科学必要性选择评测路径,进而造成"最优"声明的不可通约。

尽管近期已有多篇具身智能综述论文,如 An 等人(2025)、Sun 等人(2024)以及 Liang 等人(2025)围绕具身智能的发展进行总结,但尚未有文章对具身智能评测的方法、数据以及挑战等进行完整的梳理。上述评测体系缺失的直接后果是"横向不可比、纵向不可加":不同团队在同一任务上报告的最优性能差异可达 10% 以上,却无法判定其来源是模型创新、数据扩充还是评测偏差;产业界在选型时面临置信真空,不得不重复投入高昂的真机试验以验证厂商声明,显著抬高了技术转化门槛。因此,本文对具身智能的能力评测方法进行全景式梳理,厘清静态数据集、仿真平台与真实环境 3 类范式各自的方法学假设、适用边界与固有局限,并在此基础上提出面向标准化与可复现的未来框架,以推动具身智能领域从"技术涌现"走向"科学共识"。

1 具身智能评测背景

1.1 具身智能发展现状

近年来,随着算法、理论和硬件的飞速发展,具身智能已被广泛认为是探索通用人工智能(artificial general intelligence, AGI)的核心路径,获得了学术界和工业界的一致共识(Lake 等, 2017)。AGI 的目标是创造出能够像人类一样思考、学习和解决各种

复杂问题的智能系统。许多研究者认为,要实现这一目标,智能体必须拥有一个物理身体,使其能够像人类一样通过感知和行动与物理世界进行深度交互。这种交互不仅是获取信息的手段,更是智能形成的源泉。通过具身交互,智能体可以学习到关于物理世界的基本规律,积累常识知识,并发展出解决新问题的能力。因此,具身智能被视为弥合当前AI与AGI之间鸿沟的关键,它迫使智能体必须处理真实世界的复杂性和不确定性,从而催生出更鲁棒、更灵活和更具适应性的智能。

具身智能的发展主要围绕两大核心任务——操作(manipulation)和导航(navigation)。早期研究聚焦于特定任务和环境下训练的策略模型,形成了以强化学习(reinforcement learning, RL)和模仿学习(imitation learning, IL)为主的学习方法。随着大语言模型(large language model, LLM)和多模态大模型(multimodal large language model, MLLM)的崛起,出现了将大模型引入具身智能的具身大模型(embodied large model, ELM)范式。具身智能任务通常可形式化为部分可观测马尔可夫决策过程(partially observable Markov decision process, POMDP)。在现实环境中,智能体无法直接访问完整的环境状态,而只能通过传感器获得部分观测,因此POMDP成为描述具身交互的自然框架。一个典型的POMDP可表示为七元组$\boldsymbol{\varphi}$,具体为

$$\boldsymbol{\varphi}=(\boldsymbol{S},\boldsymbol{O},\boldsymbol{A},\boldsymbol{P},\boldsymbol{\Omega},\boldsymbol{R},\gamma) \tag{1}$$

式中,$\boldsymbol{S}$表示环境状态空间,$\boldsymbol{O}$表示观测空间,$\boldsymbol{A}$表示动作空间,$\boldsymbol{P}(\boldsymbol{S}'|\boldsymbol{S},\boldsymbol{A})$为状态转移概率分布,$\boldsymbol{\Omega}(\boldsymbol{O}|\boldsymbol{S})$为观测概率分布,$\boldsymbol{R}(\boldsymbol{S},\boldsymbol{A})$为奖励函数,$\gamma\in\boldsymbol{R}$为折扣因子。智能体的行为由策略$\pi(a|s)$控制,目标是最大化期望(max E)收益$r$,具体为

$$r=\max_{\pi}\mathrm{E}_{\pi}\left[\sum_{t}\gamma^{t}\boldsymbol{R}_{t}\right] \tag{2}$$

对于一般的具身智能任务,需要将POMDP的策略扩展为$\pi(\boldsymbol{A}|\boldsymbol{S},g)$。$g$为智能体当前的任务目标,可以由文本或数值表示。在真实环境中,状态转移$\boldsymbol{P}$和观测分布$\boldsymbol{\Omega}$较为复杂,通常不可知。仿真器(simulator)通过建立运动模型和观测模型得到近似的状态转移和观测分布。奖励函数$\boldsymbol{R}$的定义通常取决于智能体使用的学习方法。

在操作任务中,智能体通过机械臂等执行机构与环境中的物体发生交互。此时,状态空间$\boldsymbol{S}$通常包含机器人自身的运动学与动力学状态、场景中物体的位姿及其属性,以及与任务目标相关的信息。在仿真器中,观测空间$\boldsymbol{O}$可以自由定义,往往取决于实验设计;而在真实环境中,根据传感器的不同,$\boldsymbol{O}$的形式包括激光雷达点云、深度相机图像、惯性测量单元读数以及力/触觉反馈等。动作空间$\boldsymbol{A}$则取决于控制层级,可为低层的位置控制、速度控制、力控制或高层的末端轨迹规划与技能调用。

对于该任务,代表性的视觉—语言—动作大模型(vision language action model, VLA)研究,主要包括Google Robotics提出的RT系列模型。RT-1(Brohan等,2023)将Transformer引入机器人控制,实现了从多模态输入到离散化动作序列的端到端映射,显著提升了跨任务泛化能力。RT-2(Zitkovich等,2023)将视觉—语言模型(vision language model, VLM)扩展至机器人任务,通过在互联网规模的视觉问答(visual question answering, VQA)数据与机器人数据上进行协同微调,实现了从网络知识到物理执行的跨域迁移。RT-X(O'Neill等,2024)则在开放数据集Open X-Embodiment上对RT-1与RT-2进行重新训练,显著提升了二者的性能。此外,RT-Trajectory(Gu等,2024)和RT-H(Belkhale等,2024)分别在输入模态和中间层架构上对已有模型进行改进。

在导航任务中,智能体通过移动执行全局路径规划与目标到达。状态空间$\boldsymbol{S}$通常由机器人在地图坐标系下的位姿、全局地图信息和任务相关信息组成。与操作任务类似,观测空间$\boldsymbol{O}$的定义同样取决于环境、传感器和实验设计。动作空间$\boldsymbol{A}$可为连续控制命令(线速度和角速度),也可为离散动作集。

对于该任务,端到端的视觉—语言—导航大模型(vision language navigation model, VLN)的代表性工作是NaVid(Zhang等,2024c)。该方法仅依赖单目RGB视频流与自然语言指令,通过大规模视觉—语言—动作预训练,在连续环境中直接输出导航动作,实现了优秀的仿真—现实迁移性能。后续工作通过引入多任务学习、强化学习微调等机制进一步提升VLN的性能和泛化能力(Zhang等,2024b;Qi等,2025)。

随着具身智能的发展与应用,模型解决以上任务的能力也迅速提升,出现了大量的相关工作。但与此同时,随着模型能力的日益增强,需要在传统

的、基于单一任务的单一评价方法外，准确刻画模型的各项能力维度，以适应新的具身智能评测需求。

1.2 具身智能维度解耦

不同于经典的大模型评测（Zhang 等，2025d，2026），仅以最终文本或标签的正确性为唯一判据，具身智能的能力评测维度如式（1）所示，必须贯穿从状态到动作的完整链路，具体包括：在感知阶段，需考察传感器对视觉、听觉和触觉等多模态信号的保真度与时空对齐误差；在认知阶段，需评估对目标识别、场景分割与语义关联的鲁棒性，尤其关注分布外物体或遮挡条件下的表征漂移；在决策阶段，需度量任务规划、因果推理与价值对齐的合理性，检验模型能否在部分可观测环境中推演出可达且安全的行为序列；在执行阶段，则需量化动作精度、能耗、碰撞率及人机交互的物理一致性，并记录因驱动器延迟、质量—惯性不匹配导致的累积误差。如图 1 所示，只有当整条闭环的各环节误差传播被逐项分解、溯源并加权后，才能判定“失败”究竟源于传感器噪声、表征歧义、策略短视还是控制偏差，从而避免将系统级失误简单归咎于“模型输出错误”这一单一归因陷阱。

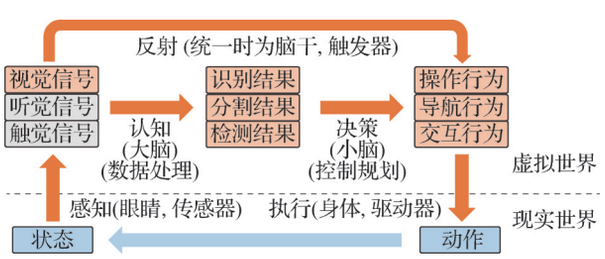

图 1 智能系统的感知—认知—决策—执行闭环

Fig. 1 Perception-cognition-decision-execution loop for intelligence systems

在百万年进化的压力下，人类已形成一套高度集成、可预测、可补偿的多层级感知—运动架构。外界光线首先经过角膜—晶状体的快速调焦和虹膜的实时光圈控制，落在具备微扫视机制的视网膜上；这些微小而高频率的眼动不仅清除了视觉暂留带来的信息衰减，还使外侧膝状体及初级视皮层能够在时序上累积高动态范围信号。随后，大脑利用自顶向下的注意信号（主要来自额叶与顶叶）对视觉场景进行软掩膜处理，抑制与当前任务无关的区域，从而节省神经计算资源。决策阶段，前额叶提供风险—收益评估，基底节与多巴胺回路给出强化信号，小脑则通过内部前馈模型预测下一步感觉输入，实现毫秒级的误差校正。最终，脊髓—肌肉—肌腱构成的可变刚度执行器，吸收突发冲击并保持动作流畅。

考虑到以上“感知—认知—决策—执行”架构的通用性，需融合认知科学的多感官加权理论，将机器智能的“信号—特征—语义—行为”与人类智能的“眼睛—大脑—小脑—身体”依次跨层耦合。现代具身智能体采用“传感器—算力—驱动器”分离的模块化架构。对于具身智能，相机以固定帧率和卷帘快门采集图像，自动曝光与白平衡算法试图模拟虹膜功能，但帧间延迟和读出时序无法用于高速运动场景；VLM 以固定感受野和位置编码提取特征，缺乏生物意义上的可塑性，需要额外注意力模块才能部分模拟额叶的“软掩膜”效果。VLN/VLA/强化学习算法通过价值函数直接输出动作分布，没有内建的前馈预测环节，必须依赖外部模型预测控制来补偿延迟，但 GPU（graphics processing unit）推理时隙限制了更新频率。执行端采用电机—减速器—连杆的高刚度链，虽然可通过力矩传感器和阻抗控制算法模拟柔性物体，但仍面临回滞与摩擦瓶颈，无法达到生物肌肉的变刚度和能量回收特性。

综上，如表 1 所示，与人类近乎理想的“感知—认知—决策—执行”耦合机制不同，现阶段的具身智能体，四大环节均存在结构性缺陷。正因如此，具身智能评测必须将四环节解耦，分别量化正确率、误差和时延等因素，才能精准定位瓶颈，为具身智能的逐层改进提供可验证的定量依据。

表 1 人类智能与具身智能的能力评测维度

Table 1 Ability evaluation dimensions for human and embodied intelligence

步骤	人类智能	具身智能
感知	角膜—晶状体—虹膜联合，微扫视清除视觉暂留	固定镜头 + 模数转换，帧间延迟导致多重失真
认知	大脑可塑性突触调整权重，抑制无关区域	VLM 固定感受野与位置编码，需额外注意力模块
决策	小脑前馈—反馈混合模型，毫秒级校正运动误差	VLA/VLN 纯价值输出，靠外部补偿延迟，受时隙限制
执行	脊髓—肌肉—肌腱可变刚度机构，提供变阻抗机制	电机—减速器—连杆高刚度链，依赖阻抗控制算法

1.3 具身智能评测范式

具身智能的评测成本与可信度呈反向指数关系:越接近真实物理世界,开销越高,但结果越可信;越依赖离线数据,开销越低,却需承受“虚实鸿沟”带来的不确定性。以下将主流评测手段归纳为三级金字塔架构——静态、仿真、真机——并在硬件预算、数据规模、部署周期与单次评估时延4个维度给出数量级对比,为研究者根据资源与精度需求选择合适范式提供量化依据。

1)静态评测。如同“笔试”,仅需GPU与若干兆字节图像或点云,无需任何机器人硬件,数分钟内即可搭建完成;一次推理耗时约0.1 s,便可将模型输出与专家标注进行开环比对。其优点是将成本压到极限,缺点是把动态交互与物理反馈完全抽象掉,模型在数据集上表现再优异,也可能因忽略几何约束或动力学延迟而在真实场景失效。

2)仿真评测。如同“面试”,需购置百元至千元级显卡与许可证,加载吉字节级的高保真场景资产,花费数小时到数天完成环境编译与接口对接;一次闭环试验约10 s,可在虚拟渲染世界里让机器人、人类与物体持续交互。由于支持物理引擎、传感器噪声和随机初始位姿,仿真能暴露部分策略短板,但仍受限于材质参数、接触模型与渲染差异,结果与真机表现间常存在可见却可控的中档落差。

3)真机评测。如同“实战”,硬件预算陡增至万元以上,数据需在现场实时采集,部署周期拉长到数天乃至数周;受限于机械拆装、标定、安全评估与人工复位,单次任务执行时间常以分钟计。其核心优势在于——物理世界不会妥协:光照变化、地面摩擦、电机回滞与意外碰撞一并涌入,任何感知漂移、预测误差或控制延迟都会立刻显形。因此,真机评测成本最高,却提供虚实差异最小的终极可信度,也是具身智能模型从实验室走向工程落地前,必须跨越的“最后一公里”。

综上,本文将依次回顾以上3种评测范式的方法,包括其任务定义、评测指标以及在具身智能“感知—认知—决策—执行”闭环中对应的步骤。

2 具身智能静态评测

2.1 评测任务

静态评测指的是在无需仿真环境或真实机器人执行的条件下,基于离线数据对具身智能模型进行能力评估的评测范式,有时也称为离线评测。

表2 具身智能评测的3种范式

Table 2 Three evaluation paradigms for embodied intelligence

资源开销	静态评测	仿真评测	真机评测
硬件费用	¥0	¥1 000～10 000	¥100 000+
数据规模	≤100 MB	1～5 GB	On-site
安装时间	分钟级	日级—周级	周级—月级
推理时间	0.1 s	10 s	60 s
可信度	低	中	高

作为“金字塔式分层架构”评价体系的第1层,静态评测旨在实现快速筛查与初步诊断。它通过低成本、高效率、可复现的离线评估过程,对具身智能模型的感知、理解与规划等核心能力进行定量测验,为后续更高成本的仿真和真实环境测试提供前置筛选与参考依据。但与此同时,静态评测与真实闭环的动态测评,也存在天然的差距。

2.2 评测指标

在具身智能领域,正确性指标最为常用。这类指标直接衡量模型是否达成任务目标,是最基础、最常用的评测方式。静态评测中常用的指标如下:

1)准确率(Acc)。即正确/错误分类样本$\boldsymbol{Class}_{\text{True}}$与$\boldsymbol{Class}_{\text{False}}$中,正确样本占总体的比例,常用于选择题类或分类任务,反映模型判断结果的整体正确性。计算为

$$Acc = \frac{\boldsymbol{Class}_{\text{True}}}{\boldsymbol{Class}_{\text{True}} + \boldsymbol{Class}_{\text{False}}} \tag{3}$$

2)MRA(mean relative accuracy)。计算为

$$MRA = \frac{1}{|\boldsymbol{C}|}\sum_{\theta \in C} 1\left(\frac{|\hat{y} - y|}{y} < 1 - \theta\right) \tag{4}$$

式中,$\hat{y}$为模型预测值,y为真实值,θ为置信阈值,集合$\boldsymbol{C} = \{0.5, 0.55, \cdots, 0.95\}$,$1(\cdot)$是指示函数,当条件成立时取1,否则取0。该式用于衡量数值预测相对准确性,平均考虑不同置信阈值下预测效果。

3)点位box的生成。即点位准确性,计算为

$$Acc_{\text{Point}} = -\frac{1}{N}\sum_{i=1}^{N} 1\left(\left\|\overline{\boldsymbol{P}_i} - \boldsymbol{P}_i\right\|_2 < \omega\right) \tag{5}$$

式中,$\overline{\boldsymbol{P}_i}$为预测点位置,$\boldsymbol{P}_i$为真实点位置,$\omega$为误差阈值,$1(\cdot)$为类似的指示函数,当预测误差小于阈值时取1,否则取0。

4)交并比定位精度。计算为

$$IoU = \frac{\boldsymbol{V}_{\text{pred}} \cap \boldsymbol{V}_{\text{gt}}}{\boldsymbol{V}_{\text{pred}} \cup \boldsymbol{V}_{\text{gt}}} \tag{6}$$

式中，$\boldsymbol{V}_{\text{pred}}$为预测的三维包围盒体积，$\boldsymbol{V}_{\text{gt}}$为真值体积。若IoU大于阈值，则视为正确定位。

5)导航误差(navigation error，NE)。为模型终点$\boldsymbol{P}_{\text{pred}}$与目标点$\boldsymbol{P}_{\text{gt}}$之间的欧氏距离。具体为

$$NE = \left\| \boldsymbol{P}_{\text{pred}} - \boldsymbol{P}_{\text{gt}} \right\|_2 \tag{7}$$

6)轨迹平均绝对误差和轨迹均方根误差。衡量模型预测轨迹$\overline{\boldsymbol{P}_t}$和真实轨迹$\boldsymbol{P}_t$在空间位置上的平均偏差。计算为

$$MSE = \frac{1}{T}\sum_{t=1}^{T} \left\| \overline{\boldsymbol{P}_t} - \boldsymbol{P}_t \right\|_1 \tag{8}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \left\| \overline{\boldsymbol{P}_t} - \boldsymbol{P}_t \right\|_2^2} \tag{9}$$

当遇到开放式问答题目的时候，一般而言，对于开放型回答的，会用计算文本相似度的指标来计算与真值标签的相似度。

7)GPT指标。计算为

$$S_{\text{GPT}} = \frac{1}{M}\sum_{i=1}^{M} GPT\left(y_i^{\text{pred}}, y_i^{\text{gt}}\right) \tag{10}$$

式中，y_i^{pred}和y_i^{gt}分别是预测和真值描述，通过$GPT(\cdot)$或类似大模型评分它们的语义一致性。

8)语义匹配度(semantic similarity)。计算为

$$S_{\text{sem}} = Sim(f_{\text{enc}}(\boldsymbol{D}_{\text{pred}}), f_{\text{enc}}(\boldsymbol{D}_{\text{gt}})) \tag{11}$$

式中，f_{enc}表示文本编码器(如CLIP、Sentence-BERT)，$\boldsymbol{D}_{\text{pred}}$和$\boldsymbol{D}_{\text{gt}}$表示全部的预测值/真值语义特征信息，$Sim(\cdot)$表示余弦相似度函数。

9)CIDEr。计算为

$$CIDEr = \frac{1}{N}\sum_{i=1}^{N} \frac{\boldsymbol{g}_i \cdot \boldsymbol{c}_i}{\left\| \boldsymbol{g}_i \right\| \left\| \boldsymbol{c}_i \right\|} \tag{12}$$

式中，$\boldsymbol{g}_i$为参考描述的特征，$\boldsymbol{c}_i$为生成描述的特征。

10)ROUGE-L。计算为

$$ROUGE = \frac{(1+\beta^2)\cdot R_{\text{LCS}} \cdot P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}} \tag{13}$$

式中，R_{LCS}为最长公共子序列在参考文本中的覆盖率，P_{LCS}为在生成文本中的覆盖率。

11)BLEU。计算为

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{14}$$

式中，p_n为n-gram精确匹配概率，w_n为权重，BP为长度惩罚。

12)TF-IDF相似度。计算为

$$Sim_{TF} = \frac{\boldsymbol{v}_a \cdot \boldsymbol{v}_b}{\left\| \boldsymbol{v}_a \right\| \left\| \boldsymbol{v}_b \right\|} \tag{15}$$

式中，$\boldsymbol{v}_a$与$\boldsymbol{v}_b$分别为生成和参考文本的TF-IDF向量。

13)文本向量余弦相似度TRF。通过编码器将文本映射到向量空间，再计算预测$\boldsymbol{f}_{\text{pred}}$与真实$\boldsymbol{f}_{\text{gt}}$向量余弦相似度，衡量语义匹配度。具体为

$$TRF = \cos\left(\boldsymbol{f}_{\text{gt}}, \boldsymbol{f}_{\text{pred}}\right) = \frac{\boldsymbol{f}_{\text{gt}} \cdot \boldsymbol{f}_{\text{pred}}}{\left\| \boldsymbol{f}_{\text{gt}} \right\| \cdot \left\| \boldsymbol{f}_{\text{pred}} \right\|} \tag{16}$$

14)效率(Efficiency)，即当前任务执行步数$Step$与理论最小需要步数$Step_{\min}$之间的比例，衡量模型在完成任务时的行为效率，值越接近1表示路径或动作规划越高效。具体为

$$Eff = \frac{Step_{\min}}{Step} \tag{17}$$

15)稳健性(PR)。该指标反映模型在随机初始化或环境扰动下性能的稳定性。计算为

$$PR = 1 - \frac{\sigma(\boldsymbol{S})}{\mu(\boldsymbol{S})} \tag{18}$$

式中，$\boldsymbol{S}$为模型在多次尝试中的成功率集合，$\mu(\cdot)$为均值，$\sigma(\cdot)$为标准差。

16)综合能力指标。该指标通过多层加权平均的方式融合不同子任务或维度的分数，反映模型多维能力综合表现。例如Uniscore，通过先计算每个原子任务o_l的得分，再对组成子任务的多个原子任务取平均得到子任务得分s_j^2，然后对组成主任务的多个子任务取平均得到主任务得分s_i^1，最后对所有主任务取平均得到总体综合得分S。即

$$S_{\text{uni}} = \frac{1}{n}\sum_{i=1}^{n} s_i^1, s_i^1 = \frac{1}{m}\sum_{j=1}^{m} s_j^2, s_j^2 = \frac{1}{k}\sum_{l=1}^{k} o_l \tag{19}$$

式中，n、m和k分别表示主任务数、子任务数和原子任务数。

此外，在实际评测中，部分工作还会对上述指标进行策略微调，如轮换评测、重复测试等，以减少随机猜测的影响并获得更加稳定可靠的结果。

2.3 评测方法

研究者提出了一系列静态评测平台和基准，用于评估具身智能模型在视觉感知、空间认知、任务推理和执行等方面的能力。它通过固定数据与环境输入，考察模型在“感知—认知—决策—执行”4个层级上的表现。相比动态闭环实验，静态评测具备高效、可复现和低风险等优势，是大规模模型筛选和对

比的重要环节(Xiao等,2025b)。如表3所示,静态　　评测在4个层级均有涉及。

表3　具身智能静态评测在感知—认知—决策—执行闭环中的应用

Table 3　Static evaluation for embodied intelligence in the perception-cognition-decision-execution loop

评测集	感知	认知	决策	执行	数据规模	核心任务	主要指标	对象	年份
Scan2Cap	√	√	–	–	51 583个描述	3D扫描场景描述	IoU, CiDEr, BLEU, METEOR, ROUGE	VLM	2021
ScanQA	√	√	–	–	41 000个问题	3D问题解答	Acc, BLEU, CIDEr ROUGE,GPT	VLM	2022
ScanRefer	√	√	–	–	51 583个描述	3D空间定位	IoU	VLM	2023
Multi3DRefer	√	√	–	–	61 900个描述	3D下的多物体定位	Recall@K, Acc, mIoU, F1-score	VLM	2023
SQA3D	√	√	–	–	20.4 k 文本描述, 33.4 k 问题	3D场景理解与推理	Acc	VLM	2023
OpenEQA	√	√	–	–	1 600+ 问题, 180+ 环境	开放词汇 EQA 数据集	Acc, GPT	VLM	2024
RoboVQA	√	√	√	–	829 502 视频文本对	长中期任务高阶推理	Acc	VLM	2024
VSI-Bench	√	√	√	–	5 000+ 问答对	视频数据集	Acc, MRA	VLM	2024
Can-Do	√	√	√	–	400 多模态样本	多样化复杂场景下具身规划任务	Acc	VLM	2024
Embodied Arena	√	√	√	√	22个Benchmark	QA/导航/任务plan	综合指标	VLM, VLN	2024
StaticEmbodiedBench	√	√	√	√	1 000条QA, 100条具身数据	机器臂操作	Acc, L2	VLM, VLA	2025
Embodied-IQA	√	√	√	√	36.9 k 条失真图像	面向机器的图像质量	Acc, BLEU, ROUGE, CIDEr	VLM, VLA	2025
ERQA	√	√	–	–	400道多选题	对物理交互至关重要的高级推理技能	Acc	VLM	2025
UniEQA	√	√	√	√	5 378问题对	理解和生成的整体基准	Uniscore, GPT	VLM	2025
VABench	√	–	–	√	300 个人工标注问题	空间可操作性,轨迹,定位box	Point Acc, MAE, RMSE, GPT	VLM, VLA	2025
PhyBlock	√	√	√		2 600个任务	QA数据集	Acc,PR	VLM	2025
MineAnyBuild	√	√	√		4 000个任务	Minecraft建筑方案	GPT	VLM	2025
OmniVTLA	√	–	–	√	40个机器人操作	私有的真实机器臂操数据集	MSE	VLA	2025
VLATest	√	–	–	√	18 604 离线测评数据	静态测试场景离线测评	MSE	VLA	2025
OpenX-Embodiment	√	–	–	√	1 M + 示例	大规模机器臂操作数据	MSE	VLA	2025
GraspVLA	√	–	–	√	10亿帧机器人抓取数据	抓取任务,提供了离线测试脚本	MSE	VLA	2025
USIM	√	–	–	√	1 852 条轨迹	水下机器臂操作任务	MSE	VLA	2025
Droid	√	–	–	√	76 K示例	大规模机器臂操作数据	MSE	VLA	2025

注:“√”表示涉及该环节,“–”表示不涉及。

感知层评测主要关注视觉属性理解、三维重建和多模态对齐能力。ScanRefer与Multi3DRefer拓展到多物体参照定位任务，采用Recall@K、mIoU和F1分数衡量多物体空间关系理解精度（Chen等，2020）。Scan2Cap首次将三维扫描场景与语言生成任务结合，通过IoU、CIDEr、BLEU和ROUGE等指标评估模型在3D视觉语义映射上的一致性（Chen等，2021）。ScanQA通过4万条问题评估3D场景理解与问答能力（Azuma等，2022）。SQA3D以33 k问题检验模型在三维空间推理中的语言—空间对齐能力（Ma等，2023）。Embodied-IQA面向具身场景图像质量评估，构建36 000多对参考与失真图像对，通过BLEU、ROUGE、CIDEr和TRF指标验证视觉信息与多模态任务的对齐一致性（Li等，2025a）。VABench在复杂空间结构下测试轨迹生成可行性，结合PointAcc与RMSE指标量化视觉表征在导航和规划中的有效性（Yuan等，2025）。

在认知层级，关注更高层次的语义和推理能力，一般以视觉问答（VQA）形式出现。Can-Do在多模态复杂场景下测试理解、推理与高阶规划能力，结合Acc和BLEU指标评估任务通用性（Chia等，2024）。RoboVQA构建大规模视频—文本问答数据集，检验模型在长时视觉推理任务中的语言生成精度（Sermanet等，2024）。OpenEQA与UniEQA代表开放词汇问答类基准，前者以Acc和GPT评估语言理解精度（Majumdar等，2024），后者通过Uniscore、Acc和BLEU综合衡量理解与生成一致性（Zhang等，2025b）。VSI-Bench聚焦视频语义整合，使用MRA和Recall@K衡量跨帧语义一致性及时空关系理解能力（Yang等，2025）。ERQA侧重物理交互因果推理，采用Acc、Recall和F1分数评估因果推理能力（Gu等，2025）。PhyBlock结合物理常识问答任务，通过Acc、precision/recall及mAP（mean average precision）衡量隐式规律抽象能力（Ma等，2025）。MineAnyBuild以Minecraft建造场景评估语言指令与生成方案一致性，采用GPT评分和指令执行匹配率（Exec Rate）量化规划与推理能力（Chen等，2025c）。Embodied Arena、StaticEmbodiedBench和UniEQA在认知层提供跨模态与多任务理解指标，从语言维度描绘具身智能系统的“语义理解—因果推理—生成一致性”链条（Ni等，2025；Xiao等，2025b；Zhang等，2024b）。

决策层级着重考察模型在任务规划与分解方面的能力，通常涉及将复杂任务拆解为有序的子任务序列。通常决策与上述认知会一同考虑。Embodied Arena整合22个评估语言理解精度（Majumdar等，2024），后者通过Uniscore、Acc和BLEU综合衡量理解与生成一致性（Ni等，2025）。StaticEmbodied Bench提供1 000条问答和100条具身操作数据，以Acc、L2距离和Task Completion Rate衡量规划合理性与指令到执行映射精度（Xiao等，2025b）。Embodied-IQA、PhyBlock及MineAnyBuild结合Acc、precision/recall和exec rate考察模型将感知与认知信息转化为可执行策略的能力（Li等，2025；Ma等，2025；Chen等，2025c）。

执行层评测着重模型动作生成与轨迹控制能力。OmniVTLA通过40条真实机器人操作轨迹测量视觉—动作对齐精度，以MSE、Trajectory Error和Success Rate量化动作误差（Cheng等，2025）。VLATest在离线仿真环境下提供18 604条轨迹，用MSE、RMSE和Success Rate评估动作生成稳定性与视觉引导精度（Wang等，2025）。GraspVLA与OpenX-Embodiment以百万级机器人示例构建跨场景抓取与操作评测体系，通过MSE、Success Rate和Grasp Acc验证执行泛化能力（Deng等，2025；O’Neiu等，2023）。USIM设计1 852条水下操作轨迹模拟低能见度与视觉扰动条件，以MSE和Task Success Rate衡量视觉—动作稳健性（Gu等，2025）。Droid通过7万多示例分析多机型、多模态对齐控制一致性，采用MSE、RMSE和Task Completion Rate指标（Khazatsky等，2024）。孟启帆等人（2025）使用虚拟现实技术在物理实验场景中评估成功率。Liu等人（2025b）则将动作空间定义为自动驾驶场景。Embodied-IQA、VABench以及StaticEmbodiedBench在执行层提供跨模态和轨迹指标，为从策略到具体动作的验证提供支持（Li等，2025a；Yuan等，2025；Xiao等，2025b）。

2.4 挑战与未来方向

总体来看，当前的静态评测方法仍然较少，主要集中在对高层视觉VLM的感知、理解与规划能力的考察，而针对具身智能中的小脑层，即VLA和VLN的控制执行的空间推理的静态评测仍相对匮乏。现有方法虽然能高效评估模型在感知和语义理解方面的表现，但在反映具身智能核心的交互性和连续控

制能力上仍存在明显不足。

首先，静态方法往往缺乏动态反馈，难以准确模拟真实具身场景中的时序变化与物理约束，从而导致与真实在线评测之间存在性能偏差。其次，各benchmark在任务设计和指标体系上缺乏统一，不同任务间难以形成可比的综合评价标准。目前常见的指标包括准确率（Acc）、均方误差（MSE）、路径偏差（RPE）、语言相关性（BLEU、ROUGE、GPT分数）等，但它们多局限于单一任务层面。

未来的研究可从3个方向推进：1）丰富静态基准的任务类型和数据多样性，使其能更全面地覆盖具身智能从感知到执行的全流程；2）探索统一的指标体系，提升跨任务与跨模型间的可比性；3）将静态与动态评测结合，通过仿真或离线交互的方式弥补两者间的差距。然而，比技术路径更关键的是范式跃迁：静态评测不应再被视为真机前的廉价筛选，而要升级为全生命周期的认知镜像。这意味着在数据生成阶段即引入物理可微渲染与神经符号混合引擎，让每一帧图像都内嵌可验证的因果链；在指标层面，用等全新测度替代单点准确率，从而提前暴露模型在十年级长周期部署中的慢性失效；最终，通过开放可扩展的“静态—仿真—真机”一体化协议栈，让全球实验室的每一次离线实验都能自动沉淀为真机系统的持续先验。

3 具身智能仿真评测

3.1 评测任务

具身智能仿真评测任务，指在具有可控物理规律、可观测状态空间及多模态交互接口的虚拟环境中，针对具身智能体的感知—认知—决策—行动全链路能力所进行的标准化、可重复的量化评测过程。其目标是：在保持环境动力学一致性与任务语义约束的前提下，通过系统化任务设计与指标体系，量化模型在多模态理解、序贯决策、物理交互与跨域泛化等方面的表现。

具身智能的仿真评测任务主要可分为三大类：视觉语言理解任务、视觉语言动作任务与视觉语言导航任务，分别面向VLM，VLA，VLN 3种模型。三者分别从语义理解、物理操作与空间决策3个维度，对具身智能体在多模态信息感知与决策执行过程中的关键能力进行标准化量化评测。

VLM评测聚焦于在感知与认知层面，对模型的多模态理解、语义对齐与视觉推理能力进行系统化评估。该类任务通常基于静态或动态视觉场景，考察模型在语言指令、视觉线索与世界知识之间的融合与推理能力。评测指标涵盖跨模态语义一致性、视觉关注与语言理解的匹配程度、视觉问答与指令理解的准确性等，旨在揭示模型在“看—懂—说”闭环中的语义对齐机制、多模态表征质量与泛化能力，为具身智能的上层认知建模提供可量化依据。

VLA评测聚焦于在具身智能体框架下，对模型的感知—理解—规划—执行全链路过程进行标准化测度。该类任务在具有动力学一致性与物理约束的仿真环境中，评估模型在基于语言指令完成物体交互与精细化操作方面的能力。评测指标涵盖视觉—语言—动作融合表征下的物理交互合理性、动作规划连续性与执行控制精度，旨在揭示模型在语义驱动的物理推理与操作决策中的有效性、可靠性与泛化规律。

VLN评测用于衡量具身智能体在自然语言指令引导下的感知、理解与序贯决策能力。该类任务通过多样化的空间场景与语言描述，量化智能体在视觉—语言—空间融合表征下的路径规划合理性、目标理解准确性与行动序列执行一致性。评测的核心在于揭示模型在语义驱动的空间推理与导航决策过程中的有效性、可靠性与泛化规律。

3.2 评测指标

仿真评测使用的大部分指标与静态评测相同。包括准确率、召回率等正确性指标，以及BLEU等语义指标。而在静态数据集之外，考虑到仿真环境中的多轮交互，其性能并不取决于二元化的成败，还需要考量如下的效率与惩罚指标：

1）任务成功分数。用于评估智能体在操作任务中达成目标状态的程度。通过计算已成功完成的原子性子任务（例如，将特定物体移动到目标位置）数量与总子任务数量的比率量化整体任务完成度。计算式为

$$S_{\text{manip}} = \frac{K'}{K} \tag{20}$$

式中，包含K个原子性子任务组成的操作任务T。K'为在任务结束时被判定为成功完成的子任务数量。

2）时间/步数（time/steps）。完成任务所需的时间或仿真步数。

3）路径长度（path length）。智能体移动的总距离。

4）交互努力（interaction effort）。如Bench-NPIN中定义的，衡量与环境交互的代价。

5）路径效率（path efficiency）。衡量智能体完成任务的导航效率，即其实际路径长度与理论最优路径长度的比值。

6）效率分数E_{manip}。该指标评估智能体在完成已成功子任务过程中的路径导航效率。它量化了智能体实际行走路径长度与完成这些子任务所需的理论最小路径长度之间的差异。

$$E_{\text{manip}} = \frac{L(K')}{l_0} \tag{21}$$

式中，$L(K')$为完成所有K'个成功子任务所需的理论最小总路径长度。这通常通过规划算法在已知环境地图上计算得出。l_0为智能体在环境中实际行走的总路径长度。

7）互动努力分数（interaction effort score）I_{nav}。该指标衡量机器人为移动环境中的物体所付出的运动学努力。它计算的是机器人移动自身所需的最小功与克服摩擦力移动所有物体（包括自身）的实际总功之间的比例。分数越高，表示不必要互动越少。计算式为

$$I_{\text{nav}} = \frac{m_0 l_0}{\sum_{i=0}^{K} m_i l_i} \tag{22}$$

式中，m_0和l_0分别是机器人的质量和移动距离。K为环境中可移动物体的总数。m_i和l_i分别是第i个可移动物体的质量和移动距离。

3.3 评测方法

本节对具身智能领域的仿真评测基准进行了系统性的梳理与剖析。如表4所示，与静态数据集不同，动态仿真很少涉及到决策与执行环节，且每种任务依赖的仿真软件不同。因此本节将现有评测基准按操作（VLA）、导航（VLN）和多模态推理（VLM）任务，即对每一种具身智能研究的核心范式依次叙述。

表4 具身智能仿真评测在感知—认知—决策—执行闭环中的应用

Table 4 Simulation evaluation for embodied intelligence in the perception-cognition-decision-execution loop

评测集	感知	认知	决策	执行	数据规模	核心任务	主要指标	对象	年份
RLBench	√	–	√	√	100个任务无限演示	视觉引导机器人操作	SR	VLA	2020
Meta-World	–	–	√	√	50个任务	多任务与元强化学习机器人操作	SR	VLA	2020
Isaac Gym	–	–	√	√	仿真平台	GPU并行机器人仿真	SR	VLA, VLN	2021
Arena-Rosnav 2.0	√	√	√	√	仿真平台	动态环境中的机器人导航	SR, Collision, Time,Path Length	VLN	2023
LIBERO	√	√	√	√	130个任务	语言条件下的终身机器人操作	SR	VLA	2023
ET-Plan-Bench	–	√	√	√	11 838个实例	低层导航、物体操作与高层任务规划	SR, Sequence Length	VLM, VLN	2024
ManiSkill3	√	–	√	√	数百万演示帧	机器人操作、移动操作与灵巧操作	SR	VLA	2024
GRUtopia	√	√	√	√	>100 000个交互场景	社交导航、物体导航与移动操作	SR	VLA, VLN	2024
GRUtopia (Loco-Navigation)	√	√	√	√	300个episodes	物体定位与移动导航	SR, Path Length	VLN	2024
GRUtopia (Loco-Manipulation)	√	√	√	√	300个episodes	移动操作	SR, Path Length	VLA	2024
RoboCasa	√	√	√	√	100个任务>10万条轨迹	厨房场景日常操作与复合任务	SR	VLA	2024

续表4　具身智能仿真评测在感知—认知—决策—执行闭环中的应用

Continued Table 4　Simulation evaluation for embodied intelligence in the perception-cognition-decision-execution loop

评测集	感知	认知	决策	执行	数据规模	核心任务	主要指标	对象	年份
GOAT-Bench	√	√	√	√	181个场景68万条episode	多模态开放词汇终身导航	SR, SPL	VLN	2024
BEDI	√	√	√	√	虚拟+真实混合评测	无人机感知、控制、工具使用与规划	SR	VLM, VLN	2025
POINT-IT-OUT	√	√	√	–	>600条人工标注	指称定位、任务驱动指向与轨迹预测	Normalized IoU	VLM	2025
Cosmos-Reason1	√	√	√	–	1 026个视频1 216个问题	物理常识、任务判断与下一动作预测	Acc	VLM	2025
ERQA (Gemini Robotics)	√	√	√	–	400个问题	具身推理视觉问答	Acc	VLM	2025
ViC-Bench	√	√	√	–	2 751个样本	迷宫、拼图、长程规划与复杂计数	Acc, Recall, Legality	VLM	2025
EgoTraj-Bench	√	√	–	–	36 947对轨迹	第一视角行人未来轨迹预测	minADE@K, minFDE@K	VLN	2025
RoboView-Bias	√	–	√	√	2 127个任务实例	机器人操作视觉偏差评测	SR, Interaction Effect	VLA	2025
RoboTwin 2.0 Benchmark	√	√	√	√	>100 000条轨迹	双臂机器人操作与泛化	SR	VLA	2025
OpenGVL	√	√	–	–	50条轨迹/数据集	时序任务进度预测	Value-Order Correlation(VOC)	VLM	2025
RoboVerse	√	–	√	√	510 500条轨迹	统一机器人操作、导航与策略学习	SR	VLA	2025
TaskSLAM-Bench	√	–	√	√	3个仿真场景2条真实路径	任务驱动SLAM与航点导航	Accuracy, Precision,Completeness	VLN	2025
HomeSafeBench	√	√	√	√	12 900条数据	自由探索式家庭安全隐患检测	Precision, Recall, F1	VLM, VLN	2025
Verti-Bench	–	–	√	√	100个环境1 000个导航任务	垂直复杂地形越野导航	SR, Traversal Time,Roll, Pitch	VLN	2025
Bench-PUSH	√	√	√	√	4类仿真任务	非抓取式交互导航与推物操作	Task Efficiency Score, Task Success Score	VLM	2025
EXPRESS-Bench	√	√	√	√	777条探索轨迹 2 044个问答–轨迹对	主动探索式具身问答	EAC	VLM	2025
Agent-RewardBench	√	√	√	–	1 136个样本	多模态智能体奖励建模	Acc	VLM	2025
SPLICE	√	√	–	–	3 381个视频11 423个事件片段	视频事件片段排序与多维推理	Binary Acc, Hamming Acc	VLM	2025

注：“√”表示涉及该环节，“–”表示不涉及。

在导航任务中，Arena-Rosnav 2.0（Kästner等，2023）是用于开发和评估机器人导航方法的平台，专注于高度动态环境中的导航任务；通过模块化设计、简化安装流程和更真实的模拟，支持多种导航方法的训练与比较，评估维度包括成功率、碰撞次数、任务完成时间及路径长度。TaskSLAM-Bench（Du等，2025）是任务驱动的SLAM（simultaneous localization and mapping）基准测试框架，用于评估SLAM方法在支持移动机器人导航至指定路径点任务时的性能，特别关注导航任务中的重复性精度，评估主要关注

精度、完整度和准确度。GOAT-Bench（Khanna等，2024）用于评估多模态终身导航能力，旨在推动通用导航模型的发展，使智能体能够无缝处理跨多种模态（物体类别、语言描述、实例图像）的目标，并在同一环境中利用过往经验进行终身学习，采用成功率及路径长度加权成功率作为主要评估指标。GRUtopia（Wang等，2024a）的基准包含明确的导航任务，如基于语言指令导航至目标物体的“Object Loco-Navigation”，以及通过与NPC交互澄清模糊指令后进行导航的“Social Loco-Navigation”，其物体定位与导航任务采用成功率、路径长度及路径长度加权成功率SPL进行评估。HomeSafeBench（Gao等，2025）用于评估具身视觉语言模型在家庭安全巡检任务中的表现；该基准通过模拟家庭环境，提供动态第一人称视角图像，支持模型在自由探索模式下识别5类常见安全隐患（火灾、触电、坠落物、绊倒、儿童安全），强调视觉感知与自主导航的融合，其主要评估指标包括精确度、召回率、F1分数及导航性能。Verti-Bench（Xu等，2025）是面向垂直挑战性非结构化地形的通用移动机器人运动能力基准测试平台，基于高保真模拟器，包含100个越野环境和1 000个导航任务，旨在客观、定量地评估和比较不同越野移动系统在极端崎岖地形下的性能，其核心评估指标为任务成功率和穿越时间。Bench-Push（Zhong等，2025）是首个针对非抓取式交互导航的综合性基准测试套件，包含一系列模拟环境，如带有可移动障碍物的迷宫导航、冰覆盖水域的自主船舶导航等，用于评估导航效率与交互努力，其迷宫导航任务以路径效率为核心指标。

对于操作任务，RLBench（James等，2020）是面向机器人操作学习的大规模基准测试与学习环境，包含100个完全独立、手工设计的任务，难度从简单的目标抓取到复杂的多阶段任务，并可通过运动规划器生成无限数量的演示轨迹，以任务成功率作为核心评估指标。

Meta-World（Yu等，2019）是用于多任务和元强化学习的开源模拟基准，提供50种不同的机器人操作任务，旨在评估算法在学习和泛化到全新操作任务上的能力，整体评估以任务成功率为核心指标。Isaac Gym（Makoviychuk等，2021）是基于GPU的高性能物理模拟平台，专为机器人学习设计，实现端到端的GPU加速训练流程，能够在单个GPU上并行运行数万个环境实例，常用于灵巧操作等复杂任务的训练与评估，在操作任务中通常以奖励和成功率作为核心评估指标。LIBERO（lifelong learning benchmark for robot manipulation）（Liu等，2023）是推动机器人操作任务中终身决策学习研究的基准，提供4个任务套件（共130个任务），强调在决策过程中对陈述性知识（物体概念）和程序性知识（动作）的混合迁移，以任务成功率作为核心评估指标。ManiSkill3（Tao等，2024）是面向通用机器人操作的高性能GPU并行化机器人仿真与渲染基准，提供涵盖移动操作、绘图、人形机器人操作、灵巧操作等12个不同领域的全面任务环境，评估主要依据任务成功率、MMRV及相关性指标。RoboCasa（Nasiriany等，2024）是以日常家庭环境（特别是厨房）为中心的大规模机器人学习基准，提供120个真实的厨房场景、100个多样化任务（原子任务和复合任务），用于训练和评估通用型机器人智能体，其原子任务以成功率为主要评估指标。RoboView-Bias（Liu等，2025a）是首个专门用于系统化量化机器人操作中视觉偏见的基准，通过在抓取任务中引入颜色、相机视角和尺度等视觉扰动因素，评估具身智能体在不同视觉条件下的性能偏差，主要评估指标包括抓取成功率、偏见系数及交互效应系数。RoboTwin 2.0（Chen等，2025b）是用于双手机器人操作的仿真框架、数据生成器和基准测试平台，旨在通过结合多模态大语言模型和仿真闭环反馈，自动生成大规模、多样化且逼真的专家数据，其内置基准测试以任务成功率为主要评价标准。RoboVerse（Geng等，2025）是统一的机器人学习基准测试框架，包含模仿学习和强化学习两大基准，并特别设计了多层次的泛化能力评估协议，核心任务围绕物体操作，其模仿学习基准以任务成功率为主要评估指标。

对于规划任务，GRUtopia（Wang等，2024a）包含复合的移动操作任务“Loco-Manipulation”，要求智能体结合移动与操作，完成“拾取—放置”任务，是导航与操作能力的综合体现；其移动操作任务采用成功率、路径长度和成功变更率进行评估。ET-Plan-Bench（Zhang等，2025a）是专门针对使用大语言模型进行具身任务规划的基准，具有一组可控且多样化的具身任务，其难度和复杂度各不相同（任务设置了成功率、动作序列长度等量化指标）。Cosmos-Reason1（Azzolini等，2025）构建了一套综合性基准，

旨在评估模型在物理常识和具身推理两方面的能力;其具身推理基准聚焦于任务完成验证、行动可行性评估和下一个合理行动预测三大关键属性,覆盖多种智能体,主要使用准确率作为评估指标。Gemini-ERQA(Abeyruwan等,2025)将Gemini的多模态推理能力扩展至物理世界,形成具身推理视觉问答能力,用于评估模型在物理世界中的时空理解与推理,采用准确率评估模型性能。PIO(Xue等,2025)是通过精确视觉基础系统评估VLM具身推理能力的基准,涵盖3个阶段(参照物体定位、任务驱动指向、视觉轨迹预测),涉及室内、厨房、驾驶和机器人操控等多个场景,使用准确率和交并比作为评估指标。ViC-Bench(Wu等,2025)专门用于评估多模态大语言模型在视觉交错思维链能力上的表现,包含4个代表性任务:迷宫导航、拼图、具身长程规划和复杂计数,要求模型基于逐步的中间视觉状态持续更新理解与决策,评估指标包括准确率、召回率和合法性。EXPRESS-Bench(Jiang等,2025)是针对探索感知的具身问答任务的大规模基准,强调智能体需在回答前主动、理性地探索环境线索,并引入探索—答案一致性指标,综合评估指标包括成功率、路径效率、测地距离和探索—答案一致性。Agent-RewardBench(Men等,2025)用于评估多模态大语言模型在智能体任务中奖励建模能力的基准,涵盖感知、规划与安全3个维度,涉及移动端、网页、自动驾驶以及虚拟家居等多种真实智能体场景,以准确率作为核心评估指标。

此外,在不涉及小脑的情况下,有关基于视觉(及多模态)信息进行理解、分析和推理的具身任务,以下评测集在特定的仿真环境中,考量了具身大脑的能力。虽然它们不属于经典的具身任务,但由于包含了与环境的交互,而非完全依赖VLM评测的静态数据集,本文仍将其视为广义的具身评测范畴。例如,CompareBench(Cai等,2025)专门用于评估视觉语言模型在视觉比较推理能力上的表现,由1 000个图像问答对组成,涵盖4个基本任务:数量比较、时间顺序、几何属性比较和空间关系推理,以准确率作为核心评估指标。Object-centric Spatial Understanding Benchmark(Mirjalili等,2025)使用合成的购物场景数据集,评估模型在3个核心任务上的表现:空间定位、空间推理(如前/后景深度排序)以及下游检索任务,专注于物体中心的细粒度空间理解。DRISHTIKON(Maji等,2025)是首个专门针对印度文化理解的多模态、多语言基准,包含64 288个图文对齐样本,覆盖丰富的文化主题,并设计了多种推理题型,用于评估模型在跨文化语境下的感知与推理能力,以准确率作为核心评估指标。SPLICE(Ballout等,2025)是基于人类标注的视觉推理评测基准,通过将视频分割为多个片段并打乱顺序,要求模型根据视觉内容重新排列这些片段以恢复原始事件顺序,评测时间、因果和空间等多维度推理能力,评估指标包括二元准确率、汉明准确率、最长公共子序列比率和编辑距离。EgoTraj-Bench(Liu等,2025c)是首个面向真实世界中第一视角噪声观测下的轨迹预测基准,旨在弥补理想化鸟瞰图轨迹预测模型与真实世界第一视角感知噪声之间的差距,推动模型在噪声环境下的鲁棒性研究,使用minADE@K和minFDE@K作为核心评估指标。是用于评估视觉语言模型在机器人任务中预测时序任务进度的开放基准,利用VLM从视觉观察中预测任务完成进度,支持零样本和少样本设置,采用值序相关性作为核心评估指标。

3.4 挑战与未来方向

当前具身智能与机器人领域的评测基准虽然已初步搭建起“感知—认知—决策—执行”全链路量测框架,但其演化轨迹仍受限于短期任务导向与仿真路径依赖。综合以上29个仿真测试集可以看出,感知与认知环节因可直接对接成熟计算机视觉与自然语言处理技术,在数据规模与指标多样性上迅速膨胀,形成“数据冗余但深度不足”的虚假繁荣;而决策与执行环节因涉及物理耦合、时序一致与安全约束,在真机环境下暴露的误差不可被静态或仿真数据稀释,导致其评测密度与迭代速度显著滞后。更关键的是,现有基准的“任务完成”定义仍停留在“单次成功”层面,忽略了长周期累积误差、非平稳环境适应与多智能体博弈等决定通用性的高阶维度,使得“高分模型”在真实部署中往往表现为“初期惊艳、后期崩溃”的脆弱曲线。

面向2030及更远的未来,具身智能评测必须从“可重复实验”升级为“可解释演化”:一方面,构建跨时间、跨空间和跨本体的“数字孪生—真机共生”闭环,让评测数据在仿真预训练、真机微调与在线更新三态之间以因果链而非快照形式持续生长,从而用误差溯源结果替代平均成功率作为模型迭代的核心

导航；另一方面，将从感知到执行的每个步骤，纳入可微分的多目标优化框架，使“性能—能效”不再是事后补丁，而是前置梯度。只有当评测基准能够前瞻性地模拟未来10年可能出现的极端场景，具身智能才有望走出实验室的理想环境，在真实世界的复杂环境、熵增中演化出可验证、可追责且可持续的通用能力。

4 真机评测

4.1 评测任务

动态真机评测指在真实物理环境中，对具身智能体进行全闭环、时序连续的测试，覆盖感知、认知、决策和执行完整链路，考核其在动态变化场景下的在线适应性与长期稳定性。

4.2 评测指标

真机评测考量的指标并不像静态与仿真评测中复杂，通常仅使用准确率与召回率等。此外，真机评测中会考量部分真实物理世界的指标，具体如下：

1）路径长度加权成功率（success weighted by path length，SPL）。该指标用于评估智能体的路径与最优路径相比的效率，同时衡量任务成功率与路径效率，避免智能体通过绕远路完成任务。计算为

$$SPL = \frac{SR \cdot L_{\text{opt}}}{\max\left(L_{\text{opt}}, L_{\text{exec}}\right)} \tag{23}$$

式中，L_{opt}表示最优路径长度，L_{exec}表示实际执行路径长度，SR为成功率。

2）执行步数Step。用以描述完成任务所需的步数，越少越高效。

3）综合评分Comp。计算为

$$Comp = \alpha_{\text{eff}} \cdot (Per + Dec) \tag{24}$$

式中，α_{eff}随步数指数衰减，Per表示感知子分，Dec表示决策子分，而感知与决策精度用各自任务下的准确率进行衡量。该框架融合了感知与决策分数。

4）相对位姿误差（relative pose error，RPE）。衡量的是估计轨迹和真值轨迹在固定间隔Δ上的局部运动（漂移）差异。

5）绝对轨迹误差（absolute trajectory error，ATE）。衡量的是估计轨迹$\boldsymbol{P}_i^{\text{gt}}$和对齐后的真值轨迹$\boldsymbol{P}_i^{\text{est}}$在全局上的差异。计算为

$$ATE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\text{trans}\left(\left(\boldsymbol{P}_i^{\text{gt}}\right)^{-1}\boldsymbol{S}\boldsymbol{P}_i^{\text{est}}\right)\right\|^2} \tag{25}$$

式中，n是位姿总数，对于相机位姿而言，需要矩阵$\boldsymbol{S} = \boldsymbol{SE}(3)$对估计位姿和参考的真实位姿进行几何变换，trans(·)表示提取平移向量。

4.3 评测方法

本节同样以“感知—认知—决策—执行”闭环为主线，将真机评测方法按主要被测环节，划分为以下4种范式，如表5所示。

感知层测试基准仅关注传感器到语义映射，而真机端用于采集或验证标注。例如对导航任务而言，只考核传感器到相机（机器人）的位姿/地图/检测的精度，不涉及指令理解与闭环控制。KITTI-odometry（Geiger等，2012）通过采集39.2 km真车序列，成为视觉—激光里程计精度应用最广泛的评价标准。Meta-World（Yu等，2019）在50个任务上真机重复试验中测量视觉编码器sim2real精度退化。HPT（heterogeneous pre-trained Transformers）（Wang等，2024b）利用多臂异构真机评估跨本体校准误差，为后续迁移策略提供感知基线。FMB（functional manipulation benchmark）（Luo等，2025）是一个功能操作真机基准，聚焦RGB-D相机在杂乱场景下的工件位姿估计。R-Bench（Li等，2025b）基于网络及轮式移动机器人采集了7类33种失真图像以及2 970对问答构建了具身视觉感知的基准。北京人形机器人公司提出了基于天工人形机器人的eSpatial-Benchmark（Zhang等，2025c），通过可配置的LEGO（leg godt）结构组装任务，严格评测模型对物理属性理解、空间依赖解析、结构稳定性推理及层级化操作序列生成的能力，耦合推理有效性于可执行动作生成中。EPD（embodied preference database）（Zhang等，2024a）提出首个面向具身图像质量感知的基准测试。EDA（embodied intelligence data assessment）（Xiao等，2025a）则通过分析1.5万条具身数据，对数据质量进行了客观化的评测。目前在真机评测方面，感知类已覆盖多模态、6D位姿和SLAM等，但跨季节、跨天气真机数据仍稀缺。

认知层测试基准关注语义到符号或关系的推理，通常不驱动执行器。例如解析语言/社交语义，决策层输出位置或路径点，侧重于语义理解、社交规则和指令解析等任务，但真机闭环仍由底层控制器

执行。NCLT（north campus long-term vision and LiDAR）（Ushani等，2015）利用15个月经过27个序列跨季节采集的基准，用于验证长期地点识别与拓扑回环检测的召回率。EQA-phys（Zhou等，2025a）提出了“物理可达性问答”真机数据集，用LLM得分量化推理正确性。CityNav（Lee等，2025）通过城市级场景采集32 637个语言目标描述和人类演示的无人机轨迹，并利用轨迹误差ATE，成功率SR，路径长度加权成功率SPL等指标衡量模型性能。上述认知类的方法虽然已经开始引入大模型零样本物理推理，但缺乏与执行闭环的深度融合。

决策层测试基准关注符号到动作序列，输出high-level指令或low-level轨迹，侧重于全局/局部路径规划、动作序列生成，但真机只做开环或少量闭环验证，一般真机端只测“规划—执行”。RLBench（James等，2020）是一个面向操作的基准和学习环境，具有100个独特的手工设计任务，旨在促进多个视觉引导操作研究领域的研究。CALVIN（composing actions from language and vision）（Mees等，2022）通过采集的1 000条真机长程轨迹，评测语言条件策略在“从语言到子目标”层面的规划成功率。RT-1/RT-2 Benchmark（Brohan等，2023；Zitkovich等，2023）通过Everyday-Robot真机采集了13万条演示，评测Transformer规划器在语言指令下的成功率。Open X-Embodiment（O’Neill等，2023）是由谷歌DeepMind联合美国斯坦福大学、上海交通大学和英伟达等21个机构，利用单臂机器人、双臂机器人和四足机器人等22种不同类型的机器采集的目前全球最大开源机器人数据集；数据集还整合了60个已有数据集，涵盖311个场景、100多万条真实机器人轨迹，包括527种技能、160 266项任务，并首次用路径长度加权成功率SPL衡量“跨本体”策略迁移的规划质量。RoboData（Yan等，2024）通过整合若干著名的数据集提供了完整的评估系统，实现了多视角图像、相机参数、深度图和动作的首次融合。RoboCerebra（Han等，2025）通过400个长程任务真机闭环，测试规划—记忆—反思能力，将执行步骤Step作为效率指标。VLABench（Zhang等，2025b）通过100个任务（包括2 000个物体），评测LLM规划器在“多步推理”场景下的长度加权成功率SPL。决策类真机评测基准呈现“语言—子目标—轨迹”统一趋势。

执行层的基准同时考核四环节，必须真机闭环。感知—认知—决策—执行四环节全部真机在线。MatterPort3D（Chang等，2017）根据Matterport下90个房间的导航任务，SPL同时考核路径效率与指令完成度。RH20T（Fang等，2024）采集了超11万帧真机长程家务操作任务的序列，用成功率考核模型的性能。ManiSkill2（Gu等，2023）利用400万个真机演示闭环，评测点云策略网络在桌面操作上的成功率。RoboMIND（Wu等，2024）是跨本体标准化的大规模数据集，包含10.7万条机器人轨迹数据，涉及479项不同任务，涵盖96种不同物体。RoboCasa（Nasiriany等，2024）通过120个任务、2 500个物体以及10万条机器人轨迹，提供“抓取和摆放，开关门，开关抽屉，转动旋钮，转动连杆，按下按钮，插入，导航”共8种基础动作来构建模块。RoboCAS（Zheng等，2024）根据杂乱场景下7 500次的真机重排，用成功率量化任务闭环精度。SimplerEnv（Li等，2025c）是sim2real对齐专用闭环基准，提出仿真和真实场景下的成功率差异指标测量模型执行性能。Fourier ActionNet（Fourier ActionNet Team和Mu，2025）采集了3万条人形灵巧双手真机演示，用成功率衡量桌面操作任务。BEDI-UAV（Guo等，2025）是面向无人机的具身评测框架，将“感知—决策—执行”解耦到标准化的子场景并给出可组合评分。AutoEval（Zhou等，2025b）提出自监督评估回路，真机闭环采集—LLM自评—更新策略，用执行成功率指标量化。RoboTwin（Chen等，2025b）包含731个已标注物体的RoboTwin-OD资产库，包含超过10万条专家轨迹的数据集。StaticEmbodiedBench（Xiao等，2025b）分别对感知—认知—决策—执行四环节进行了评测。上海人工智能实验室（Intern Robotics Team，2025）搭建了具身智能各角度的测试基准，包括操作、导航、人形以及世界模型等。执行类已走向全闭环和长程等多目标评测。

具身智能真机评测在感知—认知—决策—执行闭环中的应用如表5所示。

4.4 挑战与未来方向

对于目前的具身智能而言，真机评测存在的最大问题仍是成本高昂。对单个样本真机实验的成本，足够验证数百条仿真，甚至数千条静态样本。在成本之外，技术层面，感知—认知—决策—执行链条已被全面覆盖，但认知层仍显单薄：EQA-phys、PhysVLM-Real仅做问答，尚未与执行闭环。真机数

表 5　具身智能真机评测在感知—认知—决策—执行闭环中的应用

Table 5　Real-world evaluation for embodied intelligence in the perception-cognition-decision-execution loop

评测集	感知	认知	决策	执行	数据规模	核心任务	主要指标	对象	年份
KITTI-odometry	√	–	–	–	22 条自动驾驶序列 39.2 km路程	视觉—激光里程计	轨迹误差 ATE，RPE	汽车	2012
NCLT	√	–	–	–	15 个月采集、27 条跨季节长时视觉 + 激光雷达序列	长期地点识别、SLAM 回环检测	召回率、ATE	地面移动巡检机器人	2015
MatterPort3D	√	–	√	–	90 个室内场景 RGB-D 扫描场景	室内环境导航	SPL	室内移动机器人	2017
RLBench	√	–	√	√	100 个任务	桌面操作/强化学习/模仿学习	SR, 步数 Step	Franka	2020
Meta-World	√	–	–	√	50种任务	多任务元学习	SR	Sawyer	2020
CALVIN	√	–	√	√	1 000条真机轨迹	语言长程操作	SR	Franka	2022
RT-1 Benchmark	√	–	√	√	13万条演示	语言操作任务	SR	Everyday-Robot	2022
Open X-Embodiment	√	–	√	√	22 种不同机器人 + 527 种技能 + 超百万条演示轨迹	跨本体操作	SR	多机异构	2023
RH20T	√	–	–	√	11 万条机器人序列 + 140个任务	机器人长程操作	SR	Flexiv机械臂 + 大华-95机械爪	2023
ManiSkill2	√	–	–	√	20个任务，400万个演示帧	桌面操作/模仿学习/强化学习	SR, Acc	Franka，ROKAE xMate3Pro	2023
RT-2 Benchmark	√	√	√	√	13万交互轨迹 + 互联网数据	语言操作任务	SR	Everyday-Robot	2023
RoboMIND	√	√	√	√	10.7 万机械臂轨迹 + 479 个任务 + 96 种不同物体	生活服务	SR	Franka, 天工人形，Magic，AgileX，UR5e	2024
RoboCasa	√	–	√	√	120 种任务 + 10 万条动作	家务操作	SR	Franka机械臂 + Omron移动底盘	2024
RoboData	√	√	√	√	7万条操作轨迹	多模态操作	SR	Franka	2024
HPT	√	–	–	√	27 个异构机器人的遥操作轨迹 + 20 万条网络数据	跨本体迁移	SR	多臂异构数据训练 + Franka 测试	2024
RoboCAS	√	–	√	√	复杂布局 7 500次操作	杂乱场景操作	SR	Franka	2024
SimplerEnv	√	–	√	√	sim2real 1 500动作序列	sim2real对齐测试	sim2real SR差异	Google Robot，WidowX	2024
BEDI-UAV	√	√	√	√	154 幅图像 + 2 740 个问题测试感知；30 幅图像 + 357 个问题测试决策	无人机感知 + 决策 + 动作	Comp, Acc, Step	无人机	2025
eSpatial-Benchmark	√	√	–	–	262条颜色 + 316条质量 + 213条相对位置 + 214条尺寸	空间关系推理	Acc, SR	天工人形机器人 + Robotiq 机械手	2025
Fourier ActionNet	√	–	–	√	3 万条真机演示	人形灵巧双手	SR	Fourier全尺寸人形机器人	2025

续表5　具身智能真机评测在感知—认知—决策—执行闭环中的应用

Continued Table 5　Real-world evaluation for embodied intelligence in the perception-cognition-decision-execution loop

评测集	感知	认知	决策	执行	数据规模	核心任务	主要指标	对象	年份
EQA-phys	√	√	–	–	200个样本和1 000个真机问答	物理可达性问答	LLM-score	UR3和XArm6机器人	2025
VLABench	√	√	√	√	100个任务+2 000个物体	长时推理操作	SR, Acc	Franka	2025
RoboCerebra	√	√	√	√	400小时轨迹的长程任务	长时序操作	Acc, SR	Franka	2025
FMB	√	–	–	√	225 00条操作轨迹	抓取放置通用操作	SR	Franka	2025
AutoEval	√	√	√	√	100小时的动作序列	自主评估策略	SR	WidowX	2025
R-Bench	√	√	√	–	采集7类共33种真实图像失真样本、2970组问答样本	多模态模型抗图像失真鲁棒性评测、位姿估计	SR	轮式移动机器人AgileX Mini	2025
EDA	√	–	–	–	1.5万条具身智能采集数据	具身原始数据集质量评估、数据可学习性筛查	数据多样性熵、可学习度评分	UR5机械臂	2025
CityNav	√	√	–	–	32637条语言导航目标描述、配套无人机实测轨迹	城市户外场景语言引导无人机导航	ATE、SR、SPL	无人机	2025
RoboTwin 2.0 Benchmark	√	–	√	√	>100 000条轨迹,731个标注物体资产库	双手机器人操作	SR	多机异构机械臂	2025
StaticEmbodiedBench	√	√	√	√	1 000条QA,100条具身数据	机器臂操作	Acc	UR5机械臂	2025

注:“√”表示涉及该环节,“–”表示未涉及。

据规模呈“头部集中”:OXE、RT-1/2、RH20T三家合计超百万条,其余bench多在1～5 000量级,中小规模真机复现仍是瓶颈。因此,未来的真机评测范式有必要继续革新,例如,系统级支持7×24 h无脚本运行后,一条轨迹采集完立即回流到中央仓库,自动触发仿真校准、策略微调并生成新的待测指令,形成“仿真预训练—真机微调—在线更新”的滚动闭环。如此,真机数据可以在社区范围内持续开源迭代,成本被成百上千的用户共同摊薄,真机评测不再是经费黑洞,而是具身智能集成评测套件中的一个常用功能。

5　安全与伦理评测

5.1　评测任务

随着具身智能在各个领域的广泛应用,其复杂度和应用场景的多样化要求评测方法不断创新和完善。传统的评测方法主要集中在任务的完成度层面,已逐渐难以满足智能体在现实世界中多维度表现的全面评估需求。为此,评测的维度逐渐从单一的任务完成度,扩展到包括智能体的安全性、场景适应及部署能力、能效等多维度的综合评估,也促使评测框架朝着更加全面、系统的方向发展。评测任务的安全性主要指智能体在执行任务过程中,能够确保不对人类、环境或其他系统造成危害的能力。具体而言,安全性评测指标应关注智能体的行为决策和执行过程中是否符合预定的安全规则,并防止潜在的危险和失误。

虽然“具身安全与伦理”这一概念尚未形成统一共识,但是大模型评测领域已提出这一想法,将安全伦理视为与模型性能同等重要的因素。近年的研究已经开始通过专门设计的任务来检验智能体在“感知、认知、决策、执行”闭环中安全性、场景适应能力、以及能效等各个维度的表现,确保智能体在多种复杂情境下能够安全、可靠、高效地执行任务。

5.2 评测方法

目前，具身安全伦理层面尚无较为完善的指标，但已有部分研究人员基于人工智能三定律—不伤害人、不伤害自己、不伤害环境，对具身智能的行为进行了约束，从而在安全可信角度对具身智能进行评测。任务层面：整体的任务成功率和安全率；规划层面：能否在任务执行之前就安全的规划任务；攻防层面：能否对抗“越狱（jailbreaking）”、能否在物理后门攻击的情况下保持安全。SafeAgentBench（Yin等，2024）提出针对具身LLM代理安全意识的系统评测框架，构建了包含750个任务的安全感知基准，涵盖10类潜在风险及3种任务层次。实验结果表明，现有具身LLM在危险任务识别与拒绝上表现不足，凸显了安全感知机制在具身智能体系中的关键性。EARBench（Zhu等，2024）提出了面向具身智能任务规划安全性的自动化评测框架EARBench，构建了多模态风险场景数据集EARDataset。实验表明，当前主流大模型普遍存在较高的风险规划率，提示具身智能系统在物理环境部署前需重点强化安全意识建模与提示对齐机制。BadVLMDriver（Ni等，2024）提出了可在现实环境中触发的物理后门攻击框架，对自动驾驶的安全性进行了系统评估，表明在常见物体出现时，模型可被诱导执行危险决策，验证了攻击高效可行且具有普适性。IS-Bench（Lu等，2025）使用共161个家务任务场景，覆盖10类家庭环境，标注388个独特安全风险，在高保真物理模拟器中构造交互式场景并采用过程导向（process-oriented）评测：在风险触发步骤前/后核验是否满足相应安全目标，从而严格判定“时序正确的风险缓解”。Safe-Mind（Chen等，2025a）系统性建模具身推理链条中的安全风险，并提出可量化的评测与缓解方案。通过事实约束、因果约束和时序约束进一步界定不同类型的潜在风险。BadRobot（Zhang等，2024d）揭示了具身大型语言模型在物理世界中的越狱风险。作者提出3类攻击：通过情境角色扮演等方式诱导模型执行恶意动作；利用模型在语言与代码输出间的对齐缺陷，诱导生成有害动作；通过语义替换实现间接越狱实验显示，即使是GPT-4级别模型，在概念欺骗攻击下仍可能执行不安全动作。该工作不仅扩展了越狱研究的范畴，也为具身智能安全评测建立了标准化指标体系。Jailbreaking LLM-Controlled Robots（Robey等，2025）系统性研究了大语言模型驱动的具身机器人在越狱攻击下的安全风险，提出了一个涵盖语言到物理执行层的安全评测框架。实验结果显示，主流具身大脑控制代理在多种越狱攻击场景下均表现出显著的安全脆弱性，且防御措施常伴随虚假拒绝与性能退化。该工作将语言安全评测拓展至具身智能执行层，为后续安全防护与多模态防御机制研究提供了系统基线与指标体系。Safe-BeAl（Huang等，2025），由评测基准SafePlan-Bench与安全对齐方法Safe-Align组成，用于系统化地度量并提升LLM驱动的具身智能体在任务规划阶段的安全性，Safe-BeAl将安全性形式化为两类约束：过程安全约束与终止安全约束。AGENTSAFE（Ying等，2025）提供了覆盖“感知—规划—执行”全链路的安全评测协议与大规模风险指令集。进一步界定安全问题的源头；其思维层防御SAFE-AUDIT展示了在不显著损害效用的情况下提升安全性的可行路径。

5.3 挑战与未来方向

现阶段具身智能系统的决策空间、行动自由度与持续运行时长均受限于能源密度、控制带宽与任务场景，其整体风险仍低于家用电气与道路车辆，尚未对人类个体或社会基础设施构成实质性威胁。然而，大模型涌现的因果推理能力、与世界模型的长期规划能力可能生成超出设计者预期且难以被现有监测手段及时拦截的动作序列，从而使风险从功能失效升级为物理伤害。为应对潜在威胁演化，需在制度层面，推动建立分级认证与责任追溯体系：对超过特定能量输出或群体协作规模的系统，强制要求提交可复现的安全分析报告，确保任何异常指令在造成实质损害前被强制中断。在系统能力持续扩展的同时，维持社会风险水平，防止具身智能由辅助工具转变为需要被动应对的物理威胁源。

6 静态—仿真—真机一致性实验

为了验证静态关键帧评测方法在预测真实机器人执行任务表现中的有效性，在UR5机械臂上进行了测试，并采用Octo模型进行任务预测与执行。实验选取了50个基础任务实例，涵盖5类基础任务类型，包括拾取、放置、按压、拉和推。实验中采用两种评分机制：真实执行评分和静态评测评分。对于真实执行评分，机器人尝试完成任务，任务在以下任一

条件下终止：任务成功完成、达到最大步数限制（50步）或触发安全停止机制（如碰撞）。每个任务实例进行了50轮测试，以获取最终状态下的误差值，并据此计算位置、旋转和夹爪开合3个维度的动态评测分数。静态评测评分则通过关键帧方法获得，即由Octo模型输出与目标状态进行比较，从而直接得到静态分数，无需真实机器人执行操作。

实验主要通过计算静态分数与真实执行分数之间的皮尔逊相关系数（Pearson linear correlation coefficient，PLCC），将其定义为“静态—动态”一致性率（static-to-dynamic rate，S2D rate），来量化静态方法与真机结果之间的线性相关性。三维度分别独立计算S2D rate，并进一步计算所有维度的平均值，用于评估整体一致性水平。实验结果如图2所示，对于位置和旋转维度，图中显示静态方法分数与动态方法分数之间的相关性（S2D比率），黑色曲线为拟合趋势线，灰色区域表示bootstrap拟合曲线的±1标准差区间。对于夹爪开合维度，性能以二分类方式可视化。实验结果表明，在位置、旋转和夹爪3个控制维度上的S2D rate分别为0.629、0.621和0.729，三维度平均值为0.66，显示静态评分与真实执行评分之间具有较高的线性相关性，表明静态评测能够在整体上较可靠地反映实际执行表现。进一步按任务类型分析发现，拾取任务的S2D rate分别为位置0.796、旋转0.333、夹爪1.000；放置任务为0.471、0.343、0.750；按压任务为0.476、0.825、1.000；拉任务为0.684、0.714、0.889；推任务为0.730、0.403、1.000。结果显示，夹爪操作维度在所有任务类型中保持较高一致性，尤其对于二值控制任务表现出高度预测能力，而位置和旋转维度在不同任务类型中存在差异，例如按压任务旋转维度一致性较高，而放置任务在位置和旋转维度上一致性较低，表明静态方法在高精度动作任务预测上仍存在一定局限。

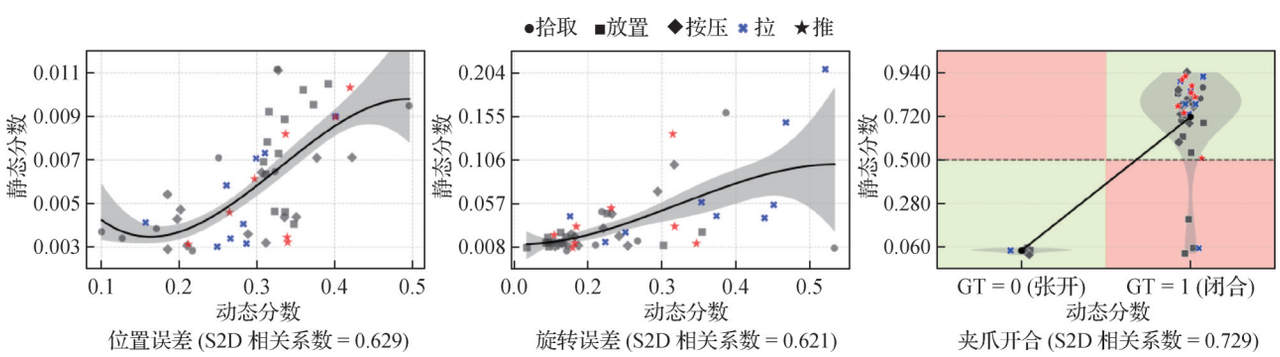

图2 Octo模型在50个任务样本上的性能散点图

Fig. 2 Scatter plot of Octo VLA model performance on 50 task samples

综合分析，本实验结果验证了静态关键帧评测方法在位置、旋转和夹爪3个关键维度上与真实执行结果具有较高一致性，表6中平均S2D rate为0.66，支持静态评测方法在大规模任务评估和模型训练中作为动态执行的代理指标的可行性。这表明，静态关键帧方法能够有效预测机器人在现实环境下的执行表现，为机器人动作模型的快速评估提供了可靠且高效的解决方案。未来，静态—仿真—真机3种评测范式将互为印证，共同构成兼顾成本与效果的评测基座方案，为具身智能的发展指明方向。

表6 不同基础任务类型在位置、旋转和夹爪开合3个维度的静态评分与真实执行评分之间的S2D系数

Table 6 The S2D correlation factor between static and real-world embodied evaluation

任务类型	位置误差	旋转误差	夹爪开合
拾取	0.796	0.333	1.000
放置	0.471	0.343	0.750
按压	0.476	0.825	1.000
拉	0.684	0.714	0.889
推	0.730	0.403	1.000

7 结 语

具身智能作为迈向通用人工智能的关键路径，其评测体系的构建已成为制约领域发展的瓶颈之一。本文围绕“从感知到执行”的完整闭环，系统梳

理了当前具身智能模型在静态数据集、仿真环境与真实物理系统3类范式下的评测方法、任务设计、指标体系与平台资源，首次从方法学假设、适用边界与固有局限3个维度对评测体系进行了全景式解构，回应了领域内“横向不可比、纵向不可加”的结构性困境。本文主要贡献与发现可归纳如下：1）体系化梳理了具身智能评测的三级范式。静态评测以低成本、高效率实现模型初筛，但脱离动态交互与物理反馈；仿真评测在可控性与真实性之间取得平衡，成为当前主流；真机评测提供最高可信度，却因成本与可复现性限制，难以规模化应用。三者构成“金字塔式”分层架构，逐级递进，互为补充。2）揭示了评测任务与指标体系的碎片化现状。当前评测任务在定义、配置与协议层面缺乏统一标准，导致同一功能在不同研究中被形式化为异质指标，模型性能差异难以归因。通过对比分析指出，任务设计需从“功能导向”转向“能力导向”，以支持对模型在感知、认知、决策、执行各环节的可解释性评估。3）提出了能力维度的解耦框架。区别于传统AI以输出正确性为唯一判据，具身智能需贯穿“感知—认知—决策—执行”全链路。本文指出，唯有将各环节误差传播逐项分解、溯源并加权，才能避免“模型输出错误”这一单一归因陷阱，为模型改进提供可验证的定量依据。

此外，本文还考量了具身智能的安全与伦理维度。随着具身智能逐步走向开放环境部署，安全评测将从“可选项”转为“必答题”。同时，识别了“仿真—现实”迁移作为核心瓶颈，如何构建可复现、可共享和可远程接入的真机评测基础设施，成为推动具身智能从“技术涌现”走向“科学共识”的关键一步。

综上所述，具身智能评测正处于从“任务驱动”向“能力导向”转型的关键阶段。未来研究需在以下方向持续深入：1）构建统一、可扩展以及可迁移的评测协议，打破平台与任务壁垒；2）推动静态—仿真—真机三级评测的协同演化，实现“仿真预训练—真机微调—在线更新”的闭环验证；3）建立安全、伦理与性能并重的多维评测框架，为具身智能的可靠部署提供制度保障。唯有如此，方能实现从“模型可比”到“能力可信”的跨越，推动具身智能走向可验证、可复现以及可落地的科学新阶段。

参考文献（References）

Abeyruwan S, Ainslie J, Alayrac J B, Arenas M G, Armstrong T, Balakrishna A, et al. 2025. Gemini robotics: bringing AI into the physical world [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2503.20020.pdf

An Z L, Yu X Q, Wang C, Zhang J L, Li Y and Song C H. 2025. Embodied intelligence: recent advances and future perspectives. The Innovation Informatics, 1(1): #100008 [DOI: 10.59717/j.xinn-inform.2025.100008]

Anderson M L. 2003. Embodied cognition: a field guide. Artificial Intelligence, 149(1): 91-130 [DOI: 10.1016/S0004-3702(03)00054-7]

Azuma D, Miyanishi T, Kurita S and Kawanabe M. 2022. Scan Q A: 3D question answering for spatial scene understanding//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19107-19117

Azzolini A, Bai J J, Brandon H, Cao J X, Chattopadhyay P, Chen H Y, et al. 2025. Cosmos-Reason1: from physical common sense to embodied reasoning [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2503.15558.pdf

Ballout M, Wilfred O, Yaghoubi S, Abdelmoneim N M, Mayer J and Bruni E. 2025. Can you SPLICE it together? A human curated benchmark for probing visual reasoning in VLMs//Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics: 11288-11309 [DOI: 10.18653/v1/2025.findings-emnlp.604]

Belkhale S, Ding T L, Xiao T, Sermanet P, Vuong Q, Tompson J, et al. 2024. RT-H: action hierarchies using language//Robotics: Science and Systems XX. Delft, the Netherlands: [s.n.]

Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, et al. 2023. RT-1: robotics transformer for real-world control at scale//Robotics: Science and Systems XIX. Daegu, Korea(South): [s.n.]

Budzianowski P, Wiśnios E, Tyrolski M, Góral G, Kulakov I, Petrenko V, et al. 2025. OpenGVL-benchmarking visual temporal progress for data curation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.17321.pdf

Cai J, Yang K N, Fu L, Ding J M, Li J L, Sun H M, et al. 2025. CompareBench: a benchmark for visual comparison reasoning in vision-language models [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.22737.pdf

Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, et al. 2017. Matterport3D: learning from RGB-D data in indoor environments//Proceedings of 2017 International Conference on 3D Vision (3DV). Qingdao, China: IEEE: 667-676 [DOI: 10.1109/3DV.2017.00081]

Chen D Z, Chang A X and Nießner M. 2020. ScanRefer: 3D object

localization in RGB-D scans using natural language//Proceedings of the 16th European Conference on Computer Vision — ECCV 2020. Glasgow, UK: Springer: 202-221 [DOI: 10.1007/978-3-030-58565-5_13]

Chen D Z, Gholami A, Nießner M and Chang A X. 2021. Scan2Cap: context-aware dense captioning in RGB-D scans//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3192-3202 [DOI: 10.1109/CVPR46437.2021.00321]

Chen R L, Sun Y Q, Wang J H, Lyu M Y, Zhang Q and Zeng Y. 2025a. SafeMind: benchmarking and mitigating safety risks in embodied LLM agents [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.25885.pdf

Chen T X, Chen Z X, Chen B J, Cai Z J, Liu Y B, Li Z X, et al. 2025b. RoboTwin 2.0: a scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2506.18088.pdf

Chen Z, Gholami A, Nießner M and Chang A X. 2025c. MineAnyBuild: benchmarking spatial planning for open-world ai agents [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2505.20148.pdf

Cheng Z X, Zhang Y Q, Zhang W K, Li H Y, Wang K Y, Song L, et al. 2025. OmniVTLA: vision-tactile-language-action model with semantic-aligned tactile sensing [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2508.08706.pdf

Chia Y K, Sun Q, Bing L D and Poria S. 2024. Can-Do! A dataset and neuro-symbolic grounded framework for embodied planning with large multimodal models [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2409.14277.pdf

Deng S L, Yan M, Wei S L, Ma H X, Yang Y X, Chen J Y, et al. 2025. GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2505.03233.pdf

Du Y W, Feng S Y, Cort C G and Vela P A. 2025. Task-driven SLAM benchmarking for robot navigation//Proceedings of 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE: 7147-7153 [DOI: 10.1109/IROS60139.2025.11246415]

Duan J F, Yu S, Tan H L, Zhu H Y and Tan C. 2022. A survey of embodied AI: from simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(2): 230-244 [DOI: 10.1109/TETCI.2022.3141105]

Fang H S, Fang H J, Tang Z Y, Liu J R, Wang C X, Wang J B, et al. 2024. RH20T: a comprehensive robotic dataset for learning diverse skills in one-shot//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE: 653-660 [DOI: 10.1109/ICRA57147.2024.10611615]

Fourier ActionNet Team and Mu Y. 2025. ActionNet: a dataset for dexterous bimanual manipulation [EB/OL]. [2025-11-05]. https://action-net.org/

Gao S Y, Yao J S, Wen H Y, Guo Y H, Liu Z M and Huang H Y. 2025. HomeSafeBench: a benchmark for embodied vision-language models in free-exploration home safety inspection [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.23690.pdf

Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 3354-3361 [DOI: 10.1109/CVPR.2012.6248074]

Geng H R, Wang F S, Wei S L, Li Y Y, Wang B J, An B S, et al. 2025. RoboVerse: towards a unified platform, dataset and benchmark for scalable and generalizable robot learning [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2504.18904.pdf

Gu J W, Wu Z H, Si P X, Qiu S, Feng Y K, Sun L Y, et al. 2025. USIM and U0: a vision-language-action dataset and model for general underwater robots [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2510.07869.pdf

Gu J Y, Kirmani S, Wohlhart P, Lu Y, Gonzalez Arenas M, Rao K, et al. 2024. RT-Trajectory: robotic task generalization via hindsight trajectory sketches//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Gu J Y, Xiang F B, Li X L, Ling Z, Liu X Q, Mu T Z, et al. 2023. ManiSkill2: a unified benchmark for generalizable manipulation skills//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Guo M N, Wu M W, He J R, Li S X, Li H F and Tao C. 2025. BEDI: a comprehensive benchmark for evaluating embodied agents on UAVs [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2505.18229.pdf

Han S H, Qiu B X, Liao Y, Huang S Y, Gao C, Yan S C, et al. 2025. RoboCerebra: a large-scale benchmark for long-horizon robotic manipulation evaluation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2506.06677.pdf

Huang Y T, Ding L L, Tang Z P, Wang T F, Lin X R, Zhang W Y, et al. 2025. A framework for benchmarking and aligning task-planning safety in LLM-based embodied agents [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2504.14650.pdf

Intern Robotics Team. 2025. InternData: a large-scale multimodal synthetic and real hybrid dataset for embodied intelligence [EB/OL]. [2025-11-05]. https://internrobotics.shlab.org.cn/opendataset.html

James S, Ma Z C, Arrojo D R and Davison A J. 2020. RLBench: the robot learning benchmark and learning environment. IEEE Robotics and Automation Letters, 5(2): 3019-3026

Jiang K X, Liu Y, Chen W X, Luo J Z, Chen Z L, Pan L, et al. 2025. Beyond the destination: a novel benchmark for exploration-aware embodied question answering [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2503.11117.pdf

Kaelbling L P, Littman M L and Moore A W. 1996. Reinforcement learn-

ing: a survey. Journal of Artificial Intelligence Research, 4(1): 237-285

Kästner L, Carstens R, Zeng H J, Kmiecik J, Bhuiyan T, Khorsandhi N, et al. 2023. Arena-Rosnav 2.0: a development and benchmarking platform for robot navigation in highly dynamic environments// Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, USA: IEEE: 11257-11264 [DOI: 10.1109/IROS55552.2023.10342152]

Khanna M, Ramrakhya R, Chhablani G, Yenamandra S, Gervet T, Chang M, et al. 2024. GOAT-Bench: a benchmark for multi-modal lifelong navigation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 16373-16383 [DOI: 10.1109/CVPR52733.2024.01549]

Khazatsky A, Pertsch K, Nair S, Balakrishna A, Dasari S, Karamcheti S, et al. 2024. DROID: a large-scale in-the-wild robot manipulation dataset//Robotics: Science and Systems XX. Delft, the Netherlands: [s.n.]

Lake B M, Ullman T D, Tenenbaum J B and Gershman S J. 2017. Building machines that learn and think like people. Behavioral and Brain Sciences, 40: #253 [DOI: 10.1017/S0140525X16001837]

LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444 [DOI: 10.1038/nature14539]

Lee J, Miyanishi T, Kurita S, Sakamoto K, Azuma D, Matsuo Y, et al. 2025. CityNav: a large-scale dataset for real-world aerial navigation//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. 5912-5922

Li C Y, Xiao J H, Zhang J B, Wen F R, Zhang Z C, Tian Y, et al. 2025a. Image quality assessment for embodied AI [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2505.16815.pdf

Li C Y, Zhang J B, Zhang Z C, Wu H N, Tian Y, Sun W, et al. 2025b. R-Bench: are your large multimodal model robust to real-world corruptions? IEEE Journal of Selected Topics in Signal Processing, 19(7): 1349-1361 [DOI: 10.1109/JSTSP.2025.3558652]

Li X, Hsu K, Gu J, Pertsch K, Mees O, Walke H R, et al. 2025c. Evaluating real-world robot manipulation policies in simulation// Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 3705-3728

Liang W L, Zhou R, Ma Y, Zhang B, Li S L, Liao Y J, et al. 2025. Large model empowered embodied AI: a survey on decision-making and embodied learning [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2508.10399.pdf

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.: #1939

Liu E G, Liang S Y, Lu L M, Zeng X Y, Cao X C, Liu A S, et al. 2025a. RoboView-Bias: benchmarking visual bias in embodied agents for robotic manipulation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.22356.pdf

Liu H R, Wan W K, Yu X Q, Li M H, Zhang J Z, Zhao B, et al. 2025b. Na Vid-4D: unleashing spatial intelligence in egocentric RGB-D videos for vision-and-language navigation//Proceedings of 2025 IEEE International Conference on Robotics and Automation. Atlanta, USA: IEEE: 10607-10615 [DOI: 10.1109/ICRA55743.2025.11128467]

Liu J Y, Zhou J M, Ye K, Lin K Y, Wang A and Liang J W. 2025c. EgoTraj-Bench: towards robust trajectory prediction under ego-view noisy observations [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2510.00405.pdf

Lu X Y, Chen Z, Hu X H, Zhou Y J, Zhang W C, Liu D R, et al. 2025. IS-Bench: evaluating interactive safety of VLM-driven embodied agents in daily household tasks [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2506.16402.pdf

Luo J L, Xu C, Liu F C, Tan L M, Lin Z P, Wu J, et al. 2025. FMB: a functional manipulation benchmark for generalizable robotic learning. The International Journal of Robotics Research, 44(4): 592-606 [DOI: 10.1177/02783649241276017]

Ma L, Wen J J, Lin M, Xu R T, Liang X W, Lin B Q, et al. 2025. PhyBlock: a progressive benchmark for physical understanding and planning via 3D block assembly [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2506.08708.pdf

Ma X J, Yong S L, Zheng Z L, Li Q, Liang Y T, Zhu S C, et al. 2023. SQA3D: situated question answering in 3D scenes//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Maji A, Kumar R, Ghosh A, Anushka, Shah N, Borah A, et al. 2025. DRISHTIKON: a multimodal multilingual benchmark for testing language models' understanding on Indian culture//Proceedings of 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics: 1289-1313 [DOI: 10.18653/v1/2025.emnlp-main.68]

Majumdar A, Ajay A, Zhang X H, Putta P, Yenamandra S, Henaff M, et al. 2024. OpenEQA: embodied question answering in the era of foundation models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 16488-16498 [DOI: 10.1109/CVPR52733.2024.01560]

Makoviychuk V, Wawrzyniak L, Guo Y R, Lu M, Storey K, Macklin, M, et al. 2021. Isaac Gym: high performance GPU-based physics simulation for robot learning//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.

Mees O, Hermann L, Rosete-Beas E and Burgard W. 2022. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters, 7(3): 7327-7334 [DOI: 10.1109/LRA.2022.3180108]

Men T Y, Jin Z R, Cao P F, Chen Y B, Liu K and Zhao J. 2025. Agent-RewardBench: towards a unified benchmark for reward modeling

across perception, planning, and safety in real-world multimodal agents//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics: 17521-17541 [DOI: 10.18653/v1/2025.acl-long.857]

Meng Q F, Zhang Y R, Zhang H W, Hu X Y and Luo Y H. 2025. Research on virtual embodied interaction technology for VR physics experiments. Journal of Image and Graphics, 1-12 (孟启帆, 张怡冉, 张鸿文, 胡晓雁, 骆岩红. 2025. 面向VR物理实验的虚拟具身交互技术研究. 中国图象图形学报, 1-12) [DOI: 10.11834/jig.250425]

Mirjalili V, Giahi R, Kollipara S, Kekuda A, Yao K H, Zhao K, et al. 2025. Spatial reasoning in foundation models: benchmarking object-centric spatial understanding [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.21922.pdf

Nasiriany S, Maddukuri A, Zhang L C, Parikh A, Lo A, Joshi A, et al. 2024. RoboCasa: large-scale simulation of everyday tasks for generalist robots [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2406.02523.pdf

Ni F, Zhang M, Li P Y, Yuan Y F, Zhang L F, Liu Y C, et al. 2025. Embodied arena: a comprehensive, unified, and evolving evaluation platform for embodied AI [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.15273.pdf

Ni Z Y, Ye R, Wei Y X, Xiang Z, Wang Y F and Chen S H. 2024. Physical backdoor attack can jeopardize driving with vision-large-language models [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2404.12916.pdf

O'Neill A, Rehman A, Gupta A, Maddukuri A, Gupta A, Padalkar A, et al. 2023. Open X-embodiment: robotic learning datasets and RT-X models [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2310.08864.pdf

O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, Lee A, et al. 2024. Open X-embodiment: robotic learning datasets and RT-X models: open X-embodiment collaboration[0]//Proceedings of 2024 IEEE International Conference on Robotics and Automation. Yokohama, Japan: IEEE: 6892-6903 [DOI: 10.1109/ICRA57147.2024.10611477]

Pfeifer R and Bongard J. 2006. How the Body Shapes the Way We Think: A New View of Intelligence. Cambridge, USA: The MIT Press

Qi Z Y, Zhang Z X, Yu Y Z, Wang J Q and Zhao H S. 2025. VLN-R1: vision-language navigation via reinforcement fine-tuning [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2506.17221.pdf

Robey A, Ravichandran Z, Kumar V, Hassani H and Pappas G J. 2025. Jailbreaking LLM-controlled robots//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 11948-11956 [DOI: 10.1109/ICRA55743.2025.11128119]

Sermanet P, Ding T L, Zhao J, Xia F, Dwibedi D, Gopalakrishnan K, et al. 2024. RoboVQA: multimodal long-horizon reasoning for robotics//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE: 645-652 [DOI: 10.1109/ICRA57147.2024.10610216]

Sun F C, Chen R F, Ji T Y, Luo Y, Zhou H D and Liu H P. 2024. A comprehensive survey on embodied intelligence: advancements, challenges, and future perspectives. CAAI Artificial Intelligence Research, 3: #9150042 [DOI: 10.26599/AIR.2024.9150042]

Tao S, Xiang F B, Shukla A, Qin Y Z, Hinrichsen X, Yuan X D, et al. 2024. ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2410.00425.pdf

Torralba A and Efros A A. 2011. Unbiased look at dataset bias//Proceedings of the CVPR 2011. Colorado Springs, USA: IEEE: 1521-1528 [DOI: 10.1109/CVPR.2011.5995347]

Ushani A K, Carlevaris-Bianco N, Cunningham A G, Galceran E and Eustice R M. 2015. Continuous-time estimation for dynamic obstacle tracking//Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany: IEEE: 1137-1143 [DOI: 10.1109/IROS.2015.7353513]

Wang H Q, Chen J H, Huang W S, Ben Q W, Wang T, Mi B Y, et al. 2024a. GRUtopia: dream general robots in a city at scale [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2407.10943.pdf

Wang L R, Chen X L, Zhao J L and He K M. 2024b. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #3952

Wang Z J, Zhou Z H, Song J Y, Huang Y H, Shu Z and Ma L. 2025. VLATest: testing and evaluating vision-language-action models for robotic manipulation. Proceedings of the ACM on Software Engineering, 2(FSE): 1615-1638 [DOI: 10.1145/3729343]

Wu K, Hou C K, Liu J M, Che Z P, Ju X Z, Yang Z Q, et al. 2024. RoboMIND: benchmark on multi-embodiment intelligence normative data for robot manipulation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2412.13877.pdf

Wu X C, Liu J X, Huang D L, Wang Y F, Shi Y Y, Chen K D, et al. 2025. ViC-bench: benchmarking visual-interleaved chain-of-thought capability in MLLMs with free-style intermediate state representations [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2505.14404.pdf

Xiao J H, Yan B W, Zhang J B, Wang J, Li C Y, Cheng Z X, et al. 2025a. Data assessment for embodied intelligence [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2508.06553.pdf

Xiao J H, Zhang J B, Yan B W, Guo S Y, Ye T R, Zhang K W, et al. 2025b. Static and plugged: make embodied evaluation simple [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2508.06553.pdf

Xu T, Pan C, Rao M B, Datar A, Pokhrel A, Lu Y, et al, 2025. Vertibench: a general and scalable off-road mobility benchmark for ver-

tically challenging terrain [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2502.11426.pdf

Xue H, Ge Y, Zeng Y, Li Z, Liu M Y, Chen Y, et al. 2025. Point-it-out: benchmarking embodied reasoning for vision language models in multi-stage visual grounding [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2509.25794.pdf

Yan F, Liu F F, Zheng L M, Zhong Y F, Huang Y Y, Guan Z C, et al. 2024. RoboTron-Mani: all-in-one multimodal large model for robotic manipulation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2412.07215.pdf

Yang J, Yang S, Gupta A W, Han R, Fei-Fei L and Xie S. 2025. Thinking in space: how multimodal large language models see, remember, and recall spaces [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2412.14171.pdf

Yin S, Pang X H, Ding Y Z, Chen M L, Bi Y T, Xiong Y C, et al. 2024. SafeAgentBench: a benchmark for safe task planning of embodied LLM agents [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2412.13178.pdf

Ying Z H, Wang L, Xiao Y S, Wang J K, Ma Y Q, Guo J Y, et al. 2025. AGENTSAFE: benchmarking the safety of embodied agents on hazardous instructions [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2506.14697.pdf

Yu T H, Quillen D, He Z P, Julian R, Hausman K, Finn C, et al. 2019. Meta-world: a benchmark and evaluation for multi-task and meta reinforcement learning//Proceedings of the 3rd Annual Conference on Robot Learning. Osaka, Japan: PMLR: 1094-1100

Yuan Y F, Cui H Q, Chen Y B, Dong Z B, Ni F, Kou L X, et al. 2025. From seeing to doing: bridging reasoning and decision for robotic manipulation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2505.08548.pdf

Zhang H, Zhu C, Wang X, Zhou Z, Hu S and Zhang L Y, 2024d. BadRobot: jailbreaking LLM-based embodied AI in the physical world [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2407.20242.pdf

Zhang J B, Li C Y, Hao J, Jia J, Duan H Y, Zheng G Q, et al. 2024a. Embodied image quality assessment for robotic intelligence [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2412.18774.pdf

Zhang J Z, Wang K Y, Wang S A, Li M H, Liu H R, Wei S L, et al. 2024b. Uni-NaVid: a video-based vision-language-action model for unifying embodied navigation tasks [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2412.06224.pdf

Zhang J Z, Wang K Y, Xu R T, Zhou G Z, Hong Y C, Fang X M, et al. 2024c. NaVid: video-based VLM plans the next step for vision-and-language navigation//Robotics: Science and Systems XX. Delft, the Netherlands: [s.n.]

Zhang L F, Wang Y N, Gu H J, Hamidizadeh A, Zhang Z G, Liu Y C, et al. 2025a. ET-plan-bench: embodied task-level planning benchmark towards spatial-temporal cognition with foundation models//Proceedings of 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE: 21566-21573 [DOI: 10.1109/IROS60139.2025.11246658]

Zhang S D, Xu Z, Liu P J, Yu X P, Li Y, Gao Q H, et al. 2025b. VLABench: a large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. [s.l.]: [s.n.]: 11142-11152

Zhang Y, Zhang Q, Ju X Z, Liu Z Y, Mao J L, Sun J K, et al. 2025c. EmbodiedVSR: dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2503.11089.pdf

Zhang Z C, Wang J Y, Guo Y J, Wen F R, Chen Z J, Wang H Q, et al. 2026. AIBench: towards trustworthy evaluation under the 45° law. Displays, 91: #103255 [DOI: 10.1016/j.displa.2025.103255]

Zheng Z C, Wang J Y, Wen F R, Guo Y J, Zhao X Y, Fang X Y, et al. 2025d. Large multimodal models evaluation: a survey. Science China Information Sciences, 68(12): #221301 [DOI: 10.1007/s11432-025-4676-4]

Zheng L M, Yan F, Liu F F, Feng C J, Kang Z L and Ma L. 2024. RoboCAS: a benchmark for robotic manipulation in complex object arrangement scenarios [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2407.06951.pdf

Zhong N H, Caro S, Iskandar A, Ramesh M and Smith S L. 2025. Bench-NPIN: benchmarking non-prehensile interactive navigation [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2505.12084.pdf

Zhou W J, Tao M L, Zhao C Y, Guo H Y, Dong H H, Tang M, et al. 2025a. PhysVLM: enabling visual language models to understand robotic physical reachability//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: 6940-6949 [DOI: 10.1109/CVPR52734.2025.00651]

Zhou Z Y, Atreya P, Tan Y L, Pertsch K and Levine S. 2025b. AutoEval: autonomous evaluation of generalist robot manipulation policies in the real world [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2503.24278.pdf

Zhu Z H, Wu B Z, Zhang Z Y, Han L, Liu Q S and Wu B Y. 2024. EARBench: towards evaluating physical risk awareness for task planning of foundation model-based embodied AI agents [EB/OL]. [2025-11-05]. https://arxiv.org/pdf/2408.04449.pdf

Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. 2023. RT-2: vision-language-action models transfer web knowledge to robotic control//Proceedings of 2023 Conference on Robot Learning. Atlanta, USA: PMLR: 2165-2183

作者简介

李春一，男，博士研究生，主要研究方向为面向具身智能的多媒体信号处理。E-mail: lichunyi@pjlab.org.cn

翟广涛，通信作者，男，教授，主要研究方向为图像处理、感知

建模、大模型与具身智能评测。
E-mail: zhaiguangtao@pjlab.org.cn
张建博，男，助理研究员，主要研究方向为具身智能大小脑协作、SLAM导航。E-mail: zhangjianbo@pjlab.org.cn
肖嘉豪，男，工程师，主要研究方向为具身智能仿真平台开发。E-mail: xiaojiahao@pjlab.org.cn
闫博闻，男，工程师，主要研究方向为具身智能仿真平台开发。E-mail: yanbowen@pjlab.org.cn
郭晟毓，男，硕士研究生，主要研究方向为具身智能与世界模型。E-mail: guoshengyu@pjlab.org.cn
叶桐瑞，男，硕士研究生，主要研究方向为具身智能心理学评测。E-mail: yetongrui@pjlab.org.cn
林维斯，男，教授，主要研究方向为图像处理、感知建模、多媒体信号处理和计算机视觉。E-mail: wslin@ntu.edu.sg