A Survey of Datasets Collection and Processing for Embodied Intelligence

DING Guiguang[1,2], ZHU Chen[1,2], WANG Xiaowan[1,2], CHEN Hui[2*]

(1. School of Software, Tsinghua University, Beijing 100084, China; 2. BNRist Tsinghua University, Beijing 100084, China)

Abstract: In recent years, vision-language-action (VLA) models have attracted significant attention in the field of embodied intelligence. As model scale continues to grow, their ability to generalize across complex tasks has steadily improved. However, such performance improvements rely heavily on the availability of large-scale, high-quality training data. Unlike natural language processing and computer vision, which can directly leverage massive internet data, data collection in embodied intelligence typically involves physical interactions between real robots and their environments, leading to high collection costs and complex acquisition processes. Efficiently obtaining, processing, and organizing such data has therefore become a critical challenge for advancing embodied intelligence. To address this issue, this paper provides a systematic review of data collection and processing methods in embodied intelligence. First, we summarize the major data acquisition paradigms from the perspective of data sources and collection strategies, and analyze their characteristics and limitations in terms of data quality, scalability, and collection cost. Second, we present a standardized processing pipeline for embodied intelligence datasets, focusing on key technical components such as action representation alignment, multimodal temporal synchronization, language semantic normalization, and data quality control. Finally, we discuss the evolving data ecosystem in embodied intelligence, highlighting current challenges and potential future directions. The analysis presented in this paper aims to provide insights for dataset construction and large-scale robot learning research in embodied intelligence.

Highlights:

1. A systematic taxonomy of embodied data acquisition paradigms is presented, analyzing the trade-offs between data quality, scalability, and collection cost across different sources.
2. A unified processing pipeline for embodied intelligence data is summarized, covering key techniques such as action representation alignment, multimodal synchronization, language normalization, and data quality control.
3. Future trends and challenges of the embodied data ecosystem are discussed, providing insights for building large-scale datasets and advancing vision-language-action based robot learning.

Key words: embodied intelligence; vision-language-action model; robot learning; large-scale data collection; data processing

---

Foundation items: National Natural Science Foundation of China (Nos.62525103, 62271281).
Received: 2026-01-09; **Revised:** 2026-02-25
***Corresponding author, E-mail:** huichen@mail.tsinghua.edu.cn.

ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 332－346
DOI:10. 16337/j. 1004-9037. 2026. 02. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

具身智能数据采集与处理综述

丁贵广[1,2]，朱　晨[1,2]，王潇婉[1,2]，陈　辉[2]

(1. 清华大学软件学院,北京 100084;2. 清华大学北京信息科学与技术国家研究中心,北京 100084)

摘　要：近年来,视觉-语言-动作(Vision-language-action,VLA)模型在具身智能领域受到广泛关注。随着模型规模不断扩大,其在复杂任务中的泛化能力持续提升,而模型性能的提升在很大程度上依赖于高质量、大规模训练数据。然而,与自然语言处理和计算机视觉领域可以直接利用互联网海量数据不同,具身智能数据通常涉及真实机器人与环境之间的物理交互,数据采集成本高、获取过程复杂。如何高效获取、处理并组织这些数据,已成为制约具身智能发展的关键问题。针对上述问题,本文对具身智能领域的数据采集与处理方法进行了系统梳理。首先,从数据来源与采集方式角度总结了当前主流的数据获取范式,并分析了不同范式在数据质量、规模潜力和采集成本等方面的特点与局限。其次,进一步总结了具身智能数据的标准化处理流程,重点分析了动作表示对齐、多模态时序同步、语言语义标准化以及数据质量控制等关键技术环节。最后,讨论了具身智能数据生态的发展趋势,指出目前遇到的困难以及未来可能的发展路径。本文的总结与分析可为具身智能领域数据集构建以及大规模机器人学习研究发展提供帮助。

关键词：具身智能;视觉-语言-动作模型;机器人学习;大规模数据采集;数据处理

中图分类号：TP18　　**文献标志码：**A

引用格式：丁贵广,朱晨,王潇婉,等.具身智能数据采集与处理综述[J].数据采集与处理,2026,41(2):332-346.
DING Guiguang,ZHU Chen,WANG Xiaowan, et al. A survey of datasets collection and processing for embodied intelligence[J]. Journal of Data Acquisition and Processing,2026,41(2):332-346.

引　　言

近年来,随着人工智能技术的快速发展,具身智能(Embodied intelligence)逐渐成为机器人研究领域的重要方向[1-2]。不同于传统仅依赖感知或决策的智能系统,具身智能强调智能体通过与真实物理世界的持续交互来学习和执行任务。在这一背景下,视觉-语言-动作(Vision-language-action,VLA)模型通过统一建模视觉感知、语言理解与动作执行,为构建通用型端到端机器人智能体提供了新的技术路径,受到学术界与工业界的广泛关注[3-4]。

现有研究表明,随着模型规模和参数量的不断扩大,VLA模型在未知任务、未知场景条件下展现出越来越强的泛化能力和鲁棒性[5-6]。这一趋势与自然语言处理和计算机视觉领域中“大模型依赖大数据”的发展规律高度一致[7-8]。然而,与这两个领域可以直接利用互联网规模的标注或弱标注数据不同,VLA模型的训练数据通常需要通过机器人与(仿真)环境的真实物理交互获得,这一过程不仅涉及多模态传感器的同步、复杂环境的构建,还需精确记录动作细节,导致数据采集成本高昂、流程繁琐,从而严重限制了高质量大规模训练数据的获取[9-10]。

基金项目：国家自然科学基金(62525103,62271281)。
收稿日期：2026-01-09;**修订日期：**2026-02-25

为降低数据采集成本并提升数据利用效率,研究者围绕具身智能数据获取开展了大量探索工作,逐步形成了多种数据采集与处理范式。例如,通过人类示教或远程操控采集高质量专家数据[11-12],利用仿真环境生成大规模合成数据[13-14],以及通过多机器人或跨平台数据共享实现数据规模扩展等[15]。这些方法在不同应用场景下各具优势,但同时各自也在数据质量、可扩展性、采集成本等方面存在一定局限性[16-17]。

此外,在具身智能领域,原始数据采集完成后的后处理环节同样至关重要。近期研究在这一方向上取得了显著进展,提出了一系列针对性的优化策略:从底层的初步采集,到中层的动作表示对齐,统一编码,多模态时序同步以及语言重标注与任务语义标准化,包括最后数据质量控制与异常检测机制[18]也被广泛引入,共同确保最终数据集的鲁棒性与泛化性。

尽管近年来相关数据集的采集与处理方法不断涌现,但目前针对具身智能数据采集与处理方法的系统性梳理仍较为缺乏。这在一定程度上增加了新研究工作的开展难度,也不利于该领域的长期可持续发展。基于此,本文围绕具身智能中的数据采集与处理问题,对现有研究工作进行系统分类与归纳,总结当前主流的数据采集范式,分析不同范式的优势与局限,对具身智能数据的处理流程进行了总结分析,并进一步探讨未来可能的发展趋势与研究方向。本文旨在从数据的视角,为后续具身智能与VLA模型相关研究提供帮助。本文核心贡献如下:

(1) 系统梳理了具身智能领域中数据采集范式,构建统一分析框架,明确不同数据来源在数据质量、规模潜力和采集成本等方面的特点与局限。

(2) 按照数据处理的流程逐步阐述各个环节的关键点与方式方法,明确每一步处理的目的与要求,并指出现存的问题与挑战。

(3) 对具身智能数据采集与处理的未来方向进行探讨,指出目前遇到的困难以及未来可能的发展路径。为构建高效、可扩展的具身智能数据体系提供参考。

1 背　　景

1.1 具身智能

具身智能强调智能体通过具备物理形态的身体与环境持续交互,在感知-决策-行动闭环中形成智能行为,被认为是通向通用人工智能的重要路径之一[2]。早期人工智能研究主要基于符号推理与规则系统构建智能模型[1],但在开放环境与复杂物理任务中表现受限。随着强化学习与机器人控制技术的发展,研究范式逐步转向数据驱动的感知-行动闭环建模,使智能体能够通过与环境交互学习控制策略。然而,纯强化学习方法往往依赖大量在线交互数据,训练成本高、样本效率低。

为缓解这一问题,模仿学习与人类示教逐渐成为重要方向,通过利用专家轨迹降低策略学习难度。近年来,随着深度学习和多模态预训练模型的快速发展,研究者开始探索将视觉、语言与动作统一建模的VLA范式[4]。代表性工作如OpenVLA[19]与π系列模型[20-22]表明,在大规模跨任务数据上进行预训练能够显著提升模型的泛化能力。这一趋势与语言模型和视觉模型中的规模法则(Scaling law)一致[7-8]。

1.2 数据集

具身智能领域中的数据并非单一模态,而是由多种信息共同构成的复杂组合。典型的具身智能数据可以表示为视觉(Vision, V_t)、语言(Language, L)、动作(Action, A_t)以及状态(State, S_t)等多模态信息的联合体。其中,视觉数据通常来自RGB或RGB-D相机,用于刻画智能体所处环境及目标对象的外观与空间关系;语言数据用于表达高层次任务目标、操作指令或人类意图;动作数据刻画机器人在连续或离散时间上的控制信号,如关节角度、关节速度或末端执行器位姿;状态数据则包含机器人自身的

本体信息与环境状态，例如关节状态、力/力矩反馈以及接触信息等[9]。

从形式化角度看，具身智能系统可以被建模为一个感知-决策-执行闭环。在时间步t时，智能体基于当前观测到的视觉信息V_t、语言指令L以及机器人状态S_t，通过策略函数π生成动作A_t，即

$$A_t=\pi(V_t,L,S_t)$$

随后，该动作在环境中执行并导致系统状态发生变化，即

$$S_{t+1}=f(S_t,A_t)$$

式中$f(\cdot)$表示环境动力学函数。由此可见，具身智能数据本质上由多模态观测与动作轨迹组成的时间序列$\{V_t|L|S_t|A_t\}_{t=1}^{T}$构成，这种多模态闭环交互数据为VLA模型学习感知-决策-动作映射关系提供了基础。

这类多模态数据在时间维度上通常呈现出显著的关联性：视觉与状态观测共同决定动作选择，而动作执行又反过来改变环境结构及后续感知结果，从而形成感知-决策-行动的闭环过程[23]。因此，与传统计算机视觉或自然语言处理中的静态样本不同，具身数据在采集过程中需要同时满足多模态同步、时序一致性以及物理可行性等约束，这也显著提高了数据采集与标注的复杂度。

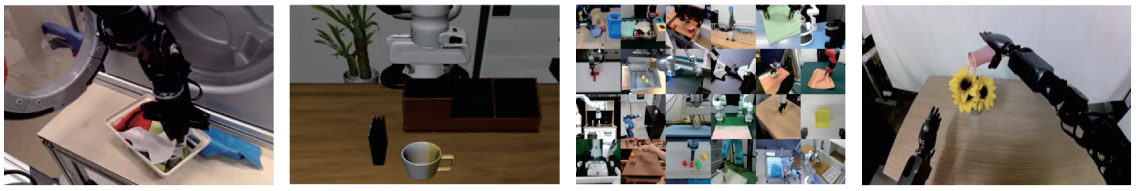
早期具身智能数据集多围绕单一或小规模任务构建，主要基于仿真环境生成状态-动作或状态-动作-奖励序列，用于强化学习与控制策略优化[23-25]。随着研究范式从强化学习逐步转向模仿学习，研究者开始系统采集真实机器人示教轨迹，构建如RoboNet[12]等大规模人类示教数据集，在提升样本效率的同时也暴露出采集成本高、扩展性受限等问题。近年来，随着通用机器人与多模态大模型的发展，具身数据集进一步向大规模、多任务与语言条件化方向演进，Meta-World[26]、LIBERO[17]等通过统一任务接口与动作空间支持多任务学习，而VLA范式则强化了视觉、语言与动作的紧密耦合。然而，由于不同机器人平台在传感器配置、执行器结构与控制接口上的差异，跨平台数据共享与复用面临显著挑战[15]，动作与状态分布的不一致加剧了标准化难度；同时，真实数据采集成本高昂，也推动了世界模型[27-28]等生成式方法在具身数据扩展中的探索与应用。

正是在上述背景下，如何高效、可扩展地采集具身智能数据，并在保证多样性与真实性的同时显著降低数据获取成本，成为当前具身智能研究中亟需解决的关键问题。后续将围绕这一核心挑战，对现有具身智能数据采集范式与处理方法进行系统梳理与分析。

2 具身智能数据采集范式

VLA模型的能力上限，在很大程度上受制于其训练数据的质量以及规模。现有研究普遍认为，单一数据来源难以支撑通用VLA模型的训练需求，因此逐渐形成了多种互补的数据采集范式。

具体而言，如图1、表1所示，现有具身智能数据采集方式可归纳为4种核心范式，分别在收集成本、数据质量以及规模方面存在显著差异，下面将对这4种范式依次进行阐述。



(a) 人类示教驱动 (a) Human demonstration-driven

(b) 仿真环境驱动 (b) Simulation environment-driven

(c) 跨平台数据聚合 (c) Cross-platform data aggregation

(d) 世界模型视频生成 (d) World model-based video generation

图1 4种具身智能数据采集方式示例

Fig.1 Examples of four data collection paradigms for embodied intelligence

表1 4种具身智能数据采集范式及其经典数据集定量定性比较

Table 1 Quantitative and qualitative comparison of four embodied intelligence data collection paradigms and representative datasets

采集方式	采集成本	数据质量	数据	轨迹数	机器人种类数	场景数目	动作种类数目	是否有语言标注
人类示教驱动	极高	高	RT-1 Datasets[5]	130 k	1	15	700	是
			BridgeData[29]	7.2 k	1	12	4	是
			BridgeData V2[30]	60.1 k	1	24	82	是
仿真环境驱动	低	一般	LIBERO[17]	6.5 k	1	5	130	是
			CALVIN[31]	约24 h	1	4	34	部分是
			Meta-World[26]	n/a	1	1	50	否
跨平台数据聚合	一般	较高	OXE[15]	1.4 M	22	311	217	部分是
			RoboTwin 2.0[32]	100 k	5	50	50	是
世界模型视频生成	低	较低	DreamGen[33]	n/a	n/a	n/a	n/a	是
			RoboEnvision[34]	n/a	n/a	n/a	n/a	是

2.1 人类示教驱动的数据采集

人类示教驱动的数据采集是具身智能领域中最早出现、也是目前应用最为成熟的数据获取方式之一。该范式通过让人类直接控制机器人执行任务,采集包含视觉观测、语言指令以及连续动作轨迹的高质量示范数据,广泛应用于模仿学习和具身智能系统中。与基于自动策略生成的数据采集方式不同,人类示教能够在数据中显式注入任务语义、动作顺序以及失败规避等高层决策信息,从而显著降低学习难度。

在具体实现上,人类示教通常采用远程操控(Teleoperation)、拖拽示教(Kinesthetic teaching)以及基于VR设备或手柄的交互控制等方式。这些方法能够将人类在操作过程中形成的连续决策过程自然映射为机器人可执行的动作序列,并在一定程度上缓解动作空间设计和奖励函数构建的困难[12]。

比如,BridgeData V2[30]数据集通过在多样化操作场景中采集人类示教轨迹,构建了大规模、多任务的操作数据集,为通用操作策略和多任务模仿学习提供了重要基础。类似地,Google提出的RT-1[5]系统通过长期的人类远程操控,在真实复杂环境中持续采集机器人操作数据,并在多种现实任务中验证了人类示教数据在大规模训练和实际部署中的可行性与稳定性。后续工作进一步表明,在此类数据基础上进行模型规模扩展,可以显著提升机器人在真实环境中的泛化能力。与自动脚本生成的仿真数据相比,基于人类示教的仿真数据在动作分布和任务执行策略上更接近真实操作习惯,因此在训练模仿学习和VLA模型时具有更好的迁移潜力[3,5]。类似的示教式仿真数据还被广泛用于早期操作学习基准中,用于验证模型在受控环境下的学习能力。

从学习范式角度看,人类示教驱动的数据通常被直接用于行为克隆(Behavior cloning)或条件模仿学习框架中,并作为VLA模型的重要监督信号[3,5]。由于示教数据隐含了较强的任务先验,模型在相对有限的数据规模下即可学到稳定且可执行的策略[35-37]。这一特性使得人类示教数据在模型初始化、技能基座构建以及关键操作能力学习阶段具有不可替代的价值。

然而,该范式的优势同时也构成了其主要局限。一方面,人类示教高度依赖人工参与,采集成本高、效率受限,难以支撑大模型训练所需的持续规模化扩展[14,30];另一方面,示教者的操作习惯、经验水平和偏好容易被模型继承,进而在面对分布外任务、不同操作风格或新机器人平台时表现出鲁棒性不

足的问题。此外,不同示教者之间的行为差异也会引入额外的数据分布不一致性,增加模型学习的难度[38-39]。

因此,人类示教驱动的数据采集通常被视为一种高质量但低扩展性的数据来源,更适合作为VLA模型的初始化数据或技能基座,而非单独支撑通用VLA模型训练的唯一数据来源。

2.2 仿真环境驱动的数据采集

仿真环境驱动的数据采集通过物理引擎构建虚拟机器人与操作场景,在无需真实硬件参与的情况下自动生成大规模交互数据,是近年来具身智能领域中发展最为迅速的一种数据采集范式[40-42]。与人类示教驱动的数据采集方式不同,该范式通常依赖程序化策略、规划器或强化学习算法在仿真环境中自主执行任务,从而自动生成包含视觉观测、状态信息与动作序列的交互数据。由于减少了人工参与,仿真驱动方法能够以较低的边际成本生成数据,并支持高度并行的数据采集流程,在数据规模和多样性方面显著优于真实示教方式。

在实际应用中,研究者通常结合领域随机化(Domain randomization)策略,对仿真环境中的视觉外观、物理参数以及初始状态进行系统性扰动,以缓解模型在真实环境中的分布偏移问题[43]。例如,在MuJoCo[44]、Isaac[45]以及Robosuite[46]等主流仿真平台中,可以随机改变物体的质量、摩擦系数、关节阻尼、接触模型以及传感器噪声分布,从而构建覆盖更广状态空间和动力学条件的数据集。该策略已被广泛用于仿真到真实(Sim-to-real)迁移研究,并在一定程度上提升了策略在真实机器人上的鲁棒性。

在此基础上,多个仿真基准被提出用于系统性评估VLA模型的能力。其中,LIBERO[17]数据集在仿真环境中通过人类演示的方式采集多任务操作轨迹,在统一的任务定义和环境配置下提供了高质量示范数据,常被用于评估模型在多任务和组合泛化设置下的性能。CALVIN[31]通过程序化生成的长时序操作任务,强调多技能的顺序组合与长期依赖建模,成为验证模型规划能力和任务组合泛化能力的重要平台。RLBench[47]则通过模块化定义的精细操作任务和自动生成的专家轨迹,提供了覆盖多种Manipulation primitive的标准化基准,广泛用于评估模型在精细操作和指令条件下的执行能力。此外,Meta-World[23,48]和ManiSkill[49]等仿真环境也通过大规模任务集合和自动化数据生成流程,为技能发现和策略泛化研究提供了重要支撑。

尽管仿真环境在视觉效果和任务多样性方面不断逼近真实世界,现有研究普遍认为Sim-to-real gap的主要瓶颈并非来自视觉逼真度不足,而是源于动作动力学和接触建模的不匹配[50-52]。即便仿真中的视觉观测高度接近真实传感器数据,动作执行过程中的微小动力学误差仍会在长时序任务中不断累积,最终导致策略在真实机器人上的性能显著退化[53-54]。针对这一问题,SIMPLER[50]等工作从系统层面对仿真数据的生成方式进行了深入分析,指出仿真数据的有效性在很大程度上取决于动作分布、控制频率以及接触动力学的一致性,而非单纯追求视觉真实性。该类研究进一步强调了在仿真数据采集过程中,对动作空间建模和动力学一致性进行约束的重要性。

从实际落地的角度来看,仿真数据采集在挖掘新技能、规划长周期任务以及搭建世界模型时,有着真实数据难以比拟的优势,毕竟在仿真里,可以较低的成本提供大量多样化的训练数据。但问题同样存在:仿真里的物理规律和动作空间建模跟现实场景中的真机实操存在差距,导致这些数据的质量远低于现实场景中采集的数据,从而影响模型训练效果。

2.3 跨平台数据聚合

近年来,机器人系统在真实场景中不断部署,逐渐存储了大量观测数据,但不同厂家所提供的机器人种类不同,其动作建模,对外接口不同,如何同时将各种不同型号的机器人数据一起用于训练成为问题。在这一背景下,跨平台数据聚合的采集方式应运而生。该范式的核心目标并非重新采集数据,而

是通过对已有异构数据进行系统性整理、对齐与统一建模,实现跨任务、跨平台的知识整合与复用。

该方向的代表性工作是Open X-Embodiment(OXE),其核心贡献并不仅体现在数据规模上,更在于对异构机器人数据的结构化统一能力。不同机器人平台在关节数量、控制频率、动作空间定义以及传感器配置等方面存在显著差异,若缺乏统一的数据抽象方式,模型难以在这些数据上进行联合训练。OXE通过引入统一的动作表示以及共享的语言语义空间,将来自不同机器人平台、不同任务设置的操作日志映射到同一表示体系中,使得VLA模型能够在多源数据上进行端到端训练。这一设计为跨机器人泛化提供了必要的结构基础。

在OXE框架的启发下,多个工作进一步探索了跨平台日志聚合在真实机器人系统中的可行性。在RoboTwin[55]的基础上,RoboTwin 2.0[32]在数据规模、任务复杂度以及机器人类型覆盖范围上进行了进一步扩展,构建了多机器人种类,特别是双臂机器人在仿真环境中的数据聚合方法。这类工作表明,只要具备合理的动作与语义对齐机制,不同机器人平台之间的数据是可以被有效整合并共同建模的。

基于上述跨平台数据聚合策略,大规模VLA模型的训练范式得以进一步扩展。例如,RT-2[4]在RT-1[5]的基础上引入了更大规模、更多来源的真实机器人操作日志,并结合视觉-语言数据进行联合训练,从而显著提升了模型在新任务和新环境中的泛化能力。OpenVLA[19]则进一步验证了在统一动作与语言表示下,通过整合多来源离线日志数据,可以训练具备较强通用操作能力的VLA模型。这些工作共同表明,多种类机器人数据在模型规模扩展和跨平台泛化中发挥着关键作用。

需要指出的是,跨平台数据聚合范式同样存在一定局限。一方面,机器人数据通常来源于既有系统的运行过程,其任务分布和动作分布难以被精确控制,可能存在覆盖不足或偏置问题;另一方面,不同平台之间的数据质量和标注规范存在差异,也对统一建模提出了更高要求。然而,与主动示教或仿真采集相比,跨平台数据聚合在规模潜力和真实物理一致性方面具有天然优势,是当前通向通用具身智能的重要数据来源之一。

2.4 基于世界模型与视频生成的具身数据合成

近年来,多个工作[18,56-60]尝试从视频-语言数据中挖掘可用于具身学习的行为结构。一些研究进一步尝试将视频-语言学习与机器人操作相结合。相关工作如VIMA[61]和BC-Z[62]等方法,尝试利用视频或图像-语言对齐模型作为中间表示,将从视频中学到的高层语义映射到机器人操作策略中。然而,仅依赖真实视频数据仍存在显著局限:视频中隐含的动作信息并未显式标注,从视觉序列恢复机器人可执行动作属于高度欠定问题,即所谓的Action recovery[63]难题。

另一方面,随着扩散模型和大规模视频生成模型的发展[64-69],例如,Video diffusion model(VDM)[70]及其后续工作展示了在长时序视频生成和复杂动态建模方面的能力,为生成具备时间一致性的操作序列提供了可能性[71]。

在此背景下,近年来逐渐出现了一类以世界模型(World model)以及视频生成模型为核心的数据合成思路[72-73],尝试通过学习环境动态和行为分布,在生成式框架下合成可用于具身智能训练的交互数据。早期工作主要聚焦于学习可预测的环境动力学模型,用于辅助规划与决策。例如,一些基于视频预测的世界模型通过在无动作或弱动作条件下建模视觉序列的演化规律,使智能体能够在潜在空间中进行多步预测,并据此进行规划或策略评估。随后,研究者开始探索将此类世界模型视为一种数据生成机制,通过在模型中采样潜在轨迹来合成新的交互序列,用于扩充训练数据或进行策略预训练。

典型代表如DreamGen[33]。该方法提出将视频世界模型作为“神经轨迹生成器(Neural trajectory generator)”,首先通过少量真实机器人示教对视频模型进行具身适配(Embodiment adaptation),使其学习特定机器人的运动学与外观特征,随后以初始帧与语言指令为条件生成大规模机器人操作视频,再

通过逆动力学模型或潜在动作模型恢复伪动作序列,最终利用生成的视频-动作对训练下游策略。

然而,基于自回归扩展的视频生成在长时序任务中容易出现误差累积与语义漂移问题。针对这一挑战,RoboEnvision[34]提出一种分层式长时序视频生成框架。该方法首先利用视觉-语言模型将高层指令分解为原子子任务,随后通过关键帧扩散模型生成与子任务对齐的阶段性关键状态,再利用插值扩散模型在关键帧之间生成连续视频,从而避免了传统自回归方式带来的长期不一致问题。

此外,也有研究将世界模型与强化学习或模仿学习框架相结合,通过在模型中进行“rollout”来生成大规模合成交互轨迹,用于策略学习或价值估计[74-75]。这类方法通常在潜在空间中进行规划与采样,再将生成结果映射回观测空间,从而形成可用于训练的伪交互数据。从范式角度看,这种方式可被视为一种“模型内仿真”,在一定程度上继承了传统仿真环境的低成本优势,同时避免了对精确物理建模的强依赖[76]。

需要指出的是,基于世界模型或视频生成的数据采集仍然面临显著挑战[77]。一方面,生成模型容易累积预测误差,尤其在长时序生成过程中可能出现物理不一致或语义漂移的问题;另一方面,生成视频中隐含的动作信息往往是隐式的,如何将其可靠地映射为机器人可执行的动作表示仍是一个开放问题。因此,当前该类方法更多被用于学习高层次的行为结构、任务先验或环境动态,而较少直接作为精细操作策略训练的唯一数据来源。

2.5 总结

综上所述,现有具身智能数据采集方法可以大致归纳为人类示教驱动、仿真环境驱动、跨平台数据聚合以及基于世界模型与视频生成的数据合成4类范式。不同数据来源在采集成本、数据质量以及规模潜力方面存在显著差异,难以由单一方式独立支撑通用VLA模型的训练需求。

从表1可以进一步观察到不同数据采集范式在数据规模和结构上的差异。人类示教数据集通常集中在单一机器人平台,例如RT-1[5]、BridgeData[29]以及BridgeData V2[30]均基于单一机器人系统采集,但在动作类型和任务语义上具有较高质量,并且普遍包含语言标注,使其在视觉-语言-动作联合学习中具有重要价值。仿真环境数据集在采集成本和任务多样性方面具有优势,例如LIBERO[17]和CALVIN[31]在统一仿真环境中构建了多个操作任务和动作原语,但其机器人类型通常仍然较为单一,这在一定程度上限制了模型在真实机器人之间的迁移能力。

相比之下,跨平台数据聚合数据集在规模和多样性方面表现出明显优势。例如OXE[15]通过整合来自22种机器人平台、311个场景的交互数据,使轨迹规模达到百万级,同时覆盖超过200种动作类型,显著提升了数据在机器人形态、任务环境和操作策略上的多样性。类似地,RoboTwin 2.0[32]也通过多机器人平台的数据整合扩大了任务覆盖范围。这类数据集的出现表明,随着VLA模型规模不断扩大,研究者越来越倾向于通过跨平台数据聚合的方式来构建更大规模、更具多样性的训练数据,从而提升模型在不同机器人平台和任务环境中的泛化能力。

此外,从表1中还可以看到近年来出现的另一类新趋势,即基于世界模型或视频生成的数据合成方法。例如DreamGen[33]和RoboEnvision[34]等工作尝试通过生成式模型合成具身交互序列,从而以较低成本扩展训练数据规模。尽管这类方法目前仍难以提供精确的动作标注,但其在语言条件、任务结构以及行为先验建模方面展现出一定潜力。

总体来看,未来VLA模型的训练可能更加依赖多来源数据的协同利用,通过结合高质量示教数据、大规模仿真数据、真实机器人日志以及生成式数据,从而在数据质量、规模和多样性之间取得更好的平衡。

3 具身智能数据处理与标准化流程

对于VLA模型而言,原始交互数据通常来源复杂,既可能包含真实机器人示教数据[25,35],也可能来自仿真环境[44,46]或跨平台大规模整合数据集[15]。这些数据在动作表示、时序频率、语言标注粒度以及数据组织结构等方面存在显著差异。若缺乏统一处理流程,模型将难以实现跨平台泛化与规模化训练。随着模型规模不断扩大[7-8],数据处理与标准化质量逐渐成为影响性能上限的重要因素。

如图2所示,具身智能数据处理通常按照由底层控制信号到高层语义信息的顺序逐步进行。首先,对不同机器人平台的控制信号进行动作表示对齐,以解决控制接口与动作空间不一致的问题,为跨平台数据融合奠定基础;随后,对视觉、语言与控制信号等多模态数据进行时序同步与轨迹重采样,以建立稳定的时间对应关系;在此基础上,对任务指令进行语言语义标准化,减少不同数据来源之间的表达差异;最后,通过数据质量控制与异常检测过滤失败轨迹与噪声样本,保证训练数据的稳定性与可靠性。通过这一由底层表示到高层语义逐层规范化的处理流程,可以提升多源具身数据的一致性与可用性。本节将对这4个环节依次阐述。

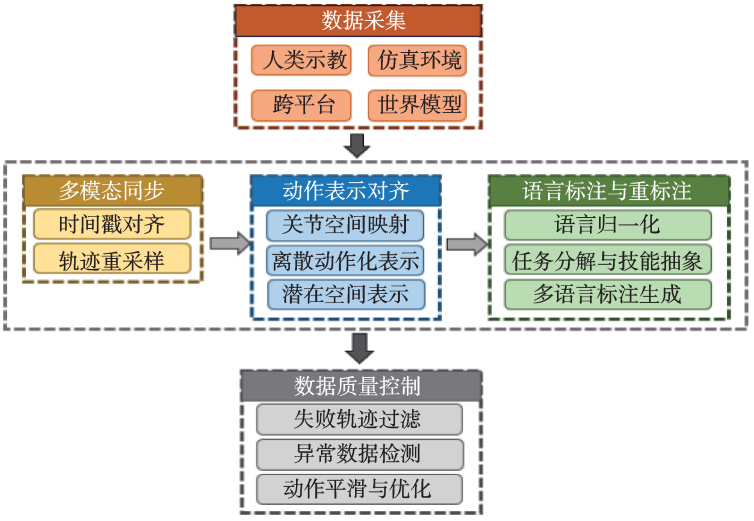


图2 具身智能数据处理流程示意图

Fig.2 Overview of the embodied intelligence data processing pipeline

3.1 动作表示对齐与统一编码

在多数据源融合背景下,动作空间不一致是首要挑战。不同机器人平台在自由度(DOF)、控制接口与动力学模型上差异显著,例如早期端到端视觉运动控制方法多采用关节空间控制[25],而大规模真实抓取系统则常基于末端执行器空间进行规划[9]。在人类示教与模仿学习框架下,动作表示往往直接继承具体平台控制接口[25,35],这在跨机器人迁移场景中会引入明显的表示偏差。

为缓解该问题,现有研究主要采用3类对齐策略。第1类为低层控制空间统一,即将不同机器人的动作映射至统一的末端执行器空间或归一化关节空间,并对幅值范围进行标准化处理,例如将关节角度或末端位姿统一映射到$[-1,1]$或$[0,1]$区间。这种方式在多机器人数据整合中较为常见,例如RoboNet[12]与OXE[15]数据整合工作均强调跨平台控制接口对齐的重要性。

第2类为离散化动作编码策略。随着VLA模型的兴起,部分方法将连续控制信号量化为离散Token序列,例如统一为$(x|y|z|r_x|r_y|r_z|g)$形式的末端执行器位姿与夹爪状态表示,将动作表示融入统

一的序列建模框架。比如RT-1[5]与RT-2[4]通过统一视觉、语言与动作表示,实现跨任务控制迁移;OpenVLA[19]与π系列模型[20-21]亦采用序列化动作建模策略以增强表示一致性。

第3类为潜在动作表示(Latent action representation)。通过自编码器或潜在规划模型学习平台无关的动作嵌入空间,可在一定程度上缓解物理接口差异带来的限制。例如从数据中学习潜在计划表示的工作[11],以及基于世界模型的潜在控制方法[28,64],均体现了将原始控制信号映射至抽象空间的趋势。

3.2 多模态时序同步与轨迹重采样

具身智能数据通常包含视觉序列、语言指令与连续控制信号3种模态,其采样频率存在显著差异。视觉帧率通常为10~30 Hz,而控制信号频率可达数十至数百赫兹[9]。语言信息则多为任务级或事件级标注[39,61]。

在真实机器人数据集中,例如BridgeData V2[30]、DROID[78]等,通常采用统一时间戳记录不同模态数据,以实现精确对齐。在仿真环境中,如MuJoCo[44]、Robosuite[46]、RLBench[47]与ManiSkill2[49],环境本身提供同步的观测与控制接口,但在导出为训练数据时仍需进行重采样与片段切分。

在实际系统中,多模态同步误差通常需要控制在10~30 ms范围内,以避免视觉观测与动作执行之间出现明显错位。在轨迹重采样过程中,常见做法是将控制信号统一重采样至10~20 Hz的训练频率。常见的处理方法包括:首先,基于时间戳的多模态对齐;其次,对高频控制信号进行下采样或对视觉帧进行插值,以匹配目标频率;再次,采用固定长度滑动窗口将长轨迹划分为可批量训练的样本。对于存在延迟与通信误差的真实机器人系统,还需通过延迟估计方法进行补偿。

随着长时序任务建模需求增强,例如CALVIN[31]等长任务基准的提出,多步时序建模对数据同步精度提出更高要求。因此,稳定且可扩展的时序对齐机制成为支撑大规模VLA训练的重要基础。

3.3 语言重标注与任务语义标准化

语言作为VLA模型的核心输入模态,在数据处理中承担语义对齐与跨任务迁移的重要角色。早期语言模仿学习工作已表明,语言表达方式的差异会显著影响模型泛化能力[39,62]。在多数据源整合背景下,同一任务可能存在不同表达方式,例如不同数据集对“抓取杯子”的描述存在词汇与句法差异。

为提高数据一致性,通常需要进行语言归一化处理,即将不同表述方式统一为标准模板,通常通过模板化指令或语义聚类实现,例如将“pick up the cup”“grab the mug”等不同表达统一映射为标准指令“grasp the cup”。

随着大语言模型的发展,研究者开始利用语言模型对原始指令进行重标注与语义增强,例如生成多种语义等价表达,以扩充训练分布。RT-2[4]与PaLM-E[6]等工作通过融合大规模视觉-语言预训练模型[60],进一步提升语言泛化能力。

此外,在复杂长任务场景下,还可将任务分解为子技能序列,并为每个子技能匹配对应语言标签。这种做法有助于建立“语言-技能-动作”三层结构,提高模型在开放场景中的可解释性与可扩展性。

3.4 数据质量控制与异常检测

原始数据中往往包含失败轨迹、异常动作与噪声样本。模仿学习与强化学习研究均指出,低质量数据可能导致策略退化与不稳定收敛[24,35]。因此,在进入训练阶段前进行数据质量控制具有必要性。

在真实机器人系统中,最常见的方法是基于任务成功信号对轨迹进行初步过滤。例如在RT-1[5]和BridgeData[29]等真实机器人数据集中,系统通常根据任务完成标志或环境状态变化判断轨迹是否成功执行,并将失败或中断的操作序列从训练数据中剔除。类似策略在RoboNet[12]等早期机器人数据集中也被广泛采用,通过自动记录任务终止状态实现对低质量轨迹的快速筛选。

除了基于任务结果的筛选,一些数据集还采用动作异常检测来识别异常样本。例如在大规模操作

数据处理中,常通过计算关节速度或控制信号的一阶、二阶差分来检测动作突变。当动作变化超过预设阈值时,该片段会被标记为异常数据并从训练集中剔除。这类方法在BridgeData[29]等操作数据处理中被用于检测传感器噪声、控制失稳或碰撞导致的异常动作。

对于跨平台聚合数据,如OXE[15],质量控制通常还需要进行数据分布过滤。由于不同实验室和机器人平台的数据在任务类型、控制频率和动作分布上存在明显差异,若直接混合训练可能导致模型过度拟合某一数据源。因此OXE[15]在数据整合过程中采用统计分析方法,对不同数据源的轨迹长度、动作幅度和任务分布进行一致性检查,并在必要时通过采样权重调整或重采样策略平衡不同数据源的比例。

总体而言,当前具身智能数据质量控制通常结合任务成功过滤、动作异常检测以及跨数据源分布平衡等多种策略,通过自动化的数据筛选流程减少噪声样本对模型训练的影响,从而提升VLA模型训练的稳定性与泛化能力。

4 未来挑战

最近,具身智能领域快速发展,相关数据采集与处理方法也快速迭代,不断优化,但具身智能数据天然存在的高成本和复杂性强等特点导致其在发展研究过程中仍面临严峻挑战。针对现有研究中的关键问题,未来可以从以下几个方向进行探索。

首先,当前主流采集方式高度依赖人类示教,数据分布往往局限于常见任务与成功轨迹,难以覆盖长尾场景与极端情况,导致模型在面对分布外任务或突发干扰时鲁棒性不足。未来可以考虑在传统示教数据的基础上引入自监督学习和元学习方法,使模型能够通过少量示教数据学习到更多的通用技能和应对突发情况的能力。

其次,为缓解真实数据采集成本高、规模受限的问题,研究者广泛引入仿真环境生成训练数据。然而,现有仿真数据在接触动力学、摩擦建模以及时序一致性等方面与真实世界仍存在明显差异,导致策略在真实机器人部署时出现性能退化。未来可以通过构建仿真-现实闭环数据生成机制来缓解这一问题,例如结合域随机化与真实数据校准不断调整仿真参数,使其更贴近真实环,逐步缩小虚实差距;另外可以探索新的训练策略,比如使用大规模仿真数据进行预训练之后利用少量高质量真实数据进行后训练微调,从而实现高规模仿真数据与真实机器人数据的协同利用。

再次,现有数据处理流程在跨平台异构数据融合方面仍存在显著瓶颈。不同机器人平台在动作空间定义、控制频率以及传感器配置等方面存在较大差异,使得跨数据源融合往往需要复杂的对齐与转换过程,容易引入噪声或丢失关键动力学信息。未来可以探索统一的动作表示与跨平台数据标准化框架,例如通过学习平台无关的潜在动作表示或技能级动作Token,将不同机器人平台的操作轨迹映射到统一表示空间。此外,可构建标准化的数据格式与接口协议,推动跨机器人平台的数据共享与互操作,从而提升大规模数据集的可扩展性。

最后,生成式人工智能与世界模型的发展为低成本数据扩展提供了新的可能,但如何将生成数据转化为可靠的具身训练数据仍是一个开放问题。当前的视频生成模型虽然能够生成视觉上逼真的序列,却难以隐式包含符合物理规律的精确动作信息,从视频中恢复可执行轨迹的“动作恢复”问题尚未得到根本解决。未来可以探索基于世界模型的行为生成框架,通过同时建模环境状态变化与动作动力学,在潜在空间中生成符合物理约束的交互轨迹。同时,可结合逆动力学模型与自动数据质量评估机制,对生成数据进行动作重建与可执行性验证,逐步构建自动化的数据生成与筛选流程,从而在保证数据质量的前提下实现具身训练数据的规模化扩展。

5 结束语

本文首先介绍了具身智能领域的发展脉络以及数据集的基本构成，然后系统梳理了具身智能领域中主流的数据采集范式，包括人类示教、仿真环境、数据聚合，以及世界生成模型合成等4类方式，并分析了各自的数据特性与适用场景。进一步分析研究了具身智能数据处理过程，包括动作表示对齐、多模态时序同步、语言语义标准化以及数据质量控制4个关键环节。最后探讨了未来具身智能数据采集的发展方向。希望能够为研究者提供一个全面了解具身智能数据采集的视角，为具身智能领域的发展做出贡献。

参考文献：

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

[2] PFEIFER R, BONGARD J. How the body shapes the way we think: A new view of intelligence[M]. [S.l.]: MIT Press, 2006.

[3] AHN M, BROHAN A, BROWN N, et al. Do as I can, not as I say: Grounding language in robotic affordances[J]. arXiv preprint arXiv: 2204.01691, 2022.

[4] ZITKOVICH B, YU T, XU S, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control[C]// Proceedings of Conference on Robot Learning. [S.l.]: PMLR, 2023: 2165-2183.

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

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

[7] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. arXiv preprint arXiv: 2001.08361, 2020.

[8] ZHAI X, KOLESNIKOV A, HOULSBY N, et al. Scaling vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 12104-12113.

[9] LEVINE S, PASTOR P, KRIZHEVSKY A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection[J]. The International Journal of Robotics Research, 2018, 37(4/5): 421-436.

[10] KROEMER O, NIEKUM S, KONIDARIS G. A review of robot learning for manipulation: Challenges, representations, and algorithms[J]. Journal of Machine Learning Research, 2021, 22(30): 1-82.

[11] LYNCH C, KHANSARI M, XIAO T, et al. Learning latent plans from play[C]//Proceedings of Conference on Robot Learning. [S.l.]: PMLR, 2020: 1113-1132.

[12] DASARI S, EBERT F, TIAN S, et al. RoboNet: Large-scale multi-robot learning[J]. arXiv preprint arXiv: 1910.11215, 2019.

[13] TOBIN J, FONG R, RAY A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C]//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [S.l.]: IEEE, 2017: 23-30.

[14] ANDRYCHOWICZ O A I M, BAKER B, CHOCIEJ M, et al. Learning dexterous in-hand manipulation[J]. The International Journal of Robotics Research, 2020, 39(1): 3-20.

[15] O'NEILL A, REHMAN A, MADDUKURI A, et al. Open X-Embodiment: Robotic learning datasets and RT-X models [C]//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2024: 6892-6903.

[16] ZHAO T Z, KUMAR V, LEVINE S, et al. Learning fine-grained bimanual manipulation with low-cost hardware[J]. arXiv preprint arXiv: 2304.13705, 2023.

[17] LIU B, ZHU Y, GAO C, et al. LIBERO: Benchmarking knowledge transfer for lifelong robot learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 44776-44791.

[18] GRAUMAN K, WESTBURY A, BYRNE E, et al. Ego4D: Around the world in 3, 000 hours of egocentric video[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 18995-19012.

[19] KIM M J, PERTSCH K, KARAMCHETI S, et al. OpenVLA: An open-source vision-language-action model[J]. arXiv preprint arXiv: 2406.09246, 2024.

[20] BLACK K, BROWN N, DRIESS D, et al. π0.5: A vision-language-action flow model for general robot control[J]. arXiv preprint arXiv: 2410.24164, 2024.

[21] BLACK K, BROWN N, DARPINIAN J, et al. π0.5: A vision-language-action model with open-world generalization[C]// Proceedings of the 9th Annual Conference on Robot Learning. [S.l.]: [s.n.], 2025.

[22] AMIN A, ANICETO R, BALAKRISHNA A, et al. π0.6: A VLA that learns from experience[J]. arXiv preprint arXiv: 2511.14759, 2025.

[23] LEVINE S, FINN C, DARRELL T, et al. End-to-end training of deep visuomotor policies[J]. Journal of Machine Learning Research, 2016, 17(39): 1-40.

[24] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: A survey[J]. The International Journal of Robotics Research, 2013, 32(11): 1238-1274.

[25] ARGALL B D, CHERNOVA S, VELOSO M, et al. A survey of robot learning from demonstration[J]. Robotics and Autonomous Systems, 2009, 57(5): 469-483.

[26] YU T, QUILLEN D, HE Z, et al. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning [C]//Proceedings of Conference on Robot Learning. [S.l.]: PMLR, 2020: 1094-1100.

[27] HA D, SCHMIDHUBER J. World models[J]. arXiv preprint arXiv: 1803.10122, 2018.

[28] HAFNER D, LILLICRAP T. Dream to control: Learning behaviors by latent imagination[C]//Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, USA: [s.n.], 2020.

[29] EBERT F, YANG Y, SCHMECKPEPER K, et al. BridgeData: Boosting generalization of robotic skills with cross-domain datasets[J]. arXiv preprint arXiv: 2109.13396, 2021.

[30] WALKE H R, BLACK K, ZHAO T Z, et al. BridgeData V2: A dataset for robot learning at scale[C]//Proceedings of Conference on Robot Learning. [S.l.]: PMLR, 2023: 1723-1736.

[31] MEES O, HERMANN L, ROSETE-BEAS E, et al. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7327-7334.

[32] CHEN T, CHEN Z, CHEN B, et al. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation[J]. arXiv preprint arXiv: 2506.18088, 2025.

[33] JANG J, YE S, LIN Z, et al. DreamGen: Unlocking generalization in robot learning through video world models[J]. arXiv preprint arXiv: 2505.12705, 2025.

[34] YANG L, BAI Y, ESKANDAR G, et al. RoboEnvision: A long-horizon video generation model for multi-task robot manipulation[J]. arXiv preprint arXiv: 2506.22007, 2025.

[35] OSA T, PAJARINEN J, NEUMANN G, et al. An algorithmic perspective on imitation learning[J]. Foundations and Trends® in Robotics, 2018, 7(1/2): 1-179.

[36] ROSS S, GORDON G, BAGNELL D. A reduction of imitation learning and structured prediction to no-regret online learning [C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. [S.l.]: JMLR Workshop and Conference Proceedings, 2011: 627-635.

[37] CALINON S. Robot programming by demonstration[M]. [S.l.]: EPFL Press, 2009.

[38] FINN C, LEVINE S, ABBEEL P. Guided cost learning: Deep inverse optimal control via policy optimization[C]// Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2016: 49-58.

[39] LYNCH C, SERMANET P. Language conditioned imitation learning over unstructured data[J]. arXiv preprint arXiv: 2005.07648, 2020.

[40] TOBIN J, FONG R, RAY A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C]//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [S.l.]: IEEE, 2017: 23-30.

[41] PENG X B, ANDRYCHOWICZ M, ZAREMBA W, et al. Sim-to-real transfer of robotic control with dynamics randomization[C]//Proceedings of 2018 IEEE International Conference on Robotics and automation (ICRA). [S.l.]: IEEE, 2018: 3803-3810.

[42] JAMES S, WOHLHART P, KALAKRISHNAN M, et al. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 12627-12637.

[43] TREMBLAY J, PRAKASH A, ACUNA D, et al. Training deep networks with synthetic data: Bridging the reality gap by domain randomization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2018: 969-977.

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

[45] MAKOVIYCHUK V, WAWRZYNIAK L, GUO Y, et al. Isaac Gym: High performance GPU-based physics simulation for robot learning[J]. arXiv preprint arXiv: 2108.10470, 2021.

[46] ZHU Y, WONG J, MANDLEKAR A, et al. Robosuite: A modular simulation framework and benchmark for robot learning [J]. arXiv preprint arXiv: 2009.12293, 2020.

[47] JAMES S, MA Z, ARROJO D R, et al. RLBench: The robot learning benchmark & learning environment[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3019-3026.

[48] MCLEAN R, CHATZAROULAS E, MCCUTCHEON L, et al. Meta-World+: An improved, standardized, RL Benchmark[J]. arXiv preprint arXiv: 2505.11289, 2025.

[49] GU J, XIANG F, LI X, et al. ManiSkill2: A unified benchmark for generalizable manipulation skills[J]. arXiv preprint arXiv: 2302.04659, 2023.

[50] LI X, HSU K, GU J, et al. Evaluating real-world robot manipulation policies in simulation[J]. arXiv preprint arXiv: 2405.05941, 2024.

[51] KADIAN A, TRUONG J, GOKASLAN A, et al. Sim2real predictivity: Does evaluation in simulation predict real-world performance?[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6670-6677.

[52] COLLINS J, HOWARD D, LEITNER J. Quantifying the reality gap in robotic manipulation tasks[C]//Proceedings of 2019 International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2019: 6706-6712.

[53] GUPTA A, MURALI A, GANDHI D P, et al. Robot learning in homes: Improving generalization and reducing dataset bias [J]. Advances in Neural Information Processing Systems, 2018, 31: 9112-9122.

[54] KALASHNIKOV D, IRPAN A, PASTOR P, et al. Scalable deep reinforcement learning for vision-based robotic manipulation[C]//Proceedings of Conference on Robot Learning. [S.l.]: PMLR, 2018: 651-673.

[55] MU Y, CHEN T, PENG S, et al. RoboTwin: Dual-arm robot benchmark with generative digital twins (early version)[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2024: 264-273.

[56] STOLFO A, BELINKOV Y, SACHAN M. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis[J]. arXiv preprint arXiv: 2305.15054, 2023.

[57] GOYAL R, EBRAHIMI KAHOU S, MICHALSKI V, et al. The "something something" video database for learning and evaluating visual common sense[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 5842-5850.

[58] DAMEN D, DOUGHTY H, FARINELLA G M, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100[J]. International Journal of Computer Vision, 2022, 130(1): 33-55.

[59] MIECH A, ZHUKOV D, ALAYRAC J B, et al. How to 100m: Learning a text-video embedding by watching hundred million narrated video clips[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 2630-2640.

[60] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.

[61] JIANG Y, GUPTA A, ZHANG Z, et al. VIMA: Robot manipulation with multimodal prompts[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 14975-15022.

[62] JANG E, IRPAN A, KHANSARI M, et al. BC-Z: Zero-shot task generalization with robotic imitation learning[C]//Proceedings of Conference on Robot Learning. [S.l.]: PMLR, 2022: 991-1002.

[63] TORABI F, WARNELL G, STONE P. Recent advances in imitation learning from observation[J]. arXiv preprint arXiv: 1905.13566, 2019.
[64] HAFNER D, LILLICRAP T, FISCHER I, et al. Learning latent dynamics for planning from pixels[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 2555-2565.
[65] HAFNER D, LILLICRAP T, NOROUZI M, et al. Mastering Atari with discrete world models[J]. arXiv preprint arXiv: 2010.02193, 2020.
[66] FINN C, TAN X Y, DUAN Y, et al. Deep spatial autoencoders for visuomotor learning[C]//Proceedings of 2016 IEEE International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2016: 512-519.
[67] OH J, GUO X, LEE H, et al. Action-conditional video prediction using deep networks in Atari games[J]. arXiv preprint arXiv: 1507.08750,2015.
[68] DENTON E, FERGUS R. Stochastic video generation with a learned prior[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2018: 1174-1183.
[69] OSINSKI B, FINN C, ERHAN D, et al. Model-based reinforcement learning for atari[J]. ICLR, 2020, 1: 2.
[70] HO J, SALIMANS T, GRITSENKO A, et al. Video diffusion models[J]. Advances in Neural Information Processing Systems, 2022, 35: 8633-8646.
[71] BLATTMANN A, ROMBACH R, LING H, et al. Align your latents: High-resolution video synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 22563-22575.
[72] QIN Y, WU Y H, LIU S, et al. DexMV: Imitation learning for dexterous manipulation from human videos[C]//Proceedings of European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 570-587.
[73] SEO Y, HAFNER D, LIU H, et al. Masked world models for visual control[C]//Proceedings of Conference on Robot Learning. [S.l.]: PMLR, 2023: 1332-1344.
[74] SUTTON R S. Dyna, an integrated architecture for learning, planning, and reacting[J]. ACM Sigart Bulletin, 1991, 2(4): 160-163.
[75] KURUTACH T, CLAVERA I, DUAN Y, et al. Model-ensemble trust-region policy optimization[J]. arXiv preprint arXiv: 1802.10592, 2018.
[76] JANNER M, FU J, ZHANG M, et al. When to trust your model: Model-based policy optimization[J]. arXiv preprint arXiv: 1906.08253,2021.
[77] TALVITIE E. Model regularization for stable sample rollouts[C]//Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. [S.l.]: [s.n.], 2014: 780-789.
[78] KHAZATSKY A, PERTSCH K, NAIR S, et al. DROID: A large-scale in-the-wild robot manipulation dataset[J]. arXiv preprint arXiv: 2403.12945, 2024.

作者简介:

丁贵广(1976-),男,教授,博士生导师,研究方向:深度学习模型设计优化、多媒体内容理解、工业视觉等,E-mail: dinggg@tsinghua.edu.cn。

朱晨(2003-),男,本科生,研究方向:具身智能数据集构建,具身智能模型的高效化微调、训练、部署以及轻量化等。

王潇婉(1989-),女,博士后,研究方向:计算传播、人工智能情感分析。

陈辉(1993-),**通信作者**,男,助理研究员,研究方向:基础视觉/多模态模型的结构设计、训练优化和推理加速等,E-mail: huichen@mail.tsinghua.edu.cn。

(编辑:夏道家)