国际先进技术应用推进中心（深圳）
CENTER FOR APPLICATION PROMOTION OF
INTERNATIONAL ADVANCED TECHNOLOGY (SHENZHEN)

具身智能数据行业研究白皮书

2026 年 3 月

目 录

表目录

图目录

01 具身智能行业的发展背景

当前，人工智能的发展处于历史性拐点，向物理世界渗透的浪潮已经展开。以大语言模型（LLM）为代表的认知智能取得巨大突破，但其能力仍主要局限于数字领域。具身智能的兴起，则致力于将智能赋予物理实体，使机器能够在现实世界中实现感知、决策与行动。

具身智能因其对多个关键领域的潜在重塑作用，战略意义显著，正迅速成为全球科技竞争的重要方向。作为实现通用人工智能（AGI）的重要路径之一，具身智能不仅引发了未来五到十年的技术创新浪潮，也有望成为推动全球生产力变革与经济增长的重要动力。

1.1 具身智能的概念与内涵

具身智能（Embodied Intelligence），作为人工智能与机器人技术交叉融合的前沿领域，其核心主张在于：高级智能的产生并非仅依赖于抽象计算过程，而是与物理实体及其所处环境之间的持续感知 – 行动循环密切相关。具身智能的概念源于认知一元论，该理论主张所有已知形式的智能，包括人类智能，本质上都是具身化的，即智能体必须具有物理形态并与环境直接交互。这一核心范式的转变，促进了数据驱动的人工智能和场景驱动的机器人深度融合。

以大语言模型为代表的认知智能进展，正与具身智能形成互补与融合。从技术架构上看，具身智能系统构成了一个由大脑（Brain）、身体（Body）与环境（Environment）三者构成的动态耦合系统。大模型为机器人提供了高层任务规划、常识推理和自然语言交互的能力，承担“认知大脑”角色；机器人本体则作为大模型在物理世界中的“行动载体”，赋予数字智能以实体化与执行能力。这种认知与身体的有机结合，推动具身智能从执行预设任务的自动化设备，向能够理解开放指令、适应非结构化环境的通用智能体发展。

1.2 具身智能正发展成为全球科技焦点

从美国将其纳入国家关键技术发展路径，到中国将其视为产业升级与培育新质生产力的关键方向；从科技企业加速整合资源、推出人形机器人原型产品，到全球资本在产业链关键环节进行前瞻性布局，围绕该领域的国际竞争正全面展开，具身智能正逐渐成为全球科技竞争的战略焦点。

从国家战略层面看，主要经济体已通过政策引导将具身智能纳入科技竞争框架。美国政府通过《国家机器人计划》等专项政策持续支持基础研究与产业应用，白宫 2025 年 7 月发布美国 AI 行动计划，指出关注人工智能在物理世界的创新，优先投资有关无人机、自动驾驶汽车、机器人等发明创造。欧盟则借助“地平线欧洲”等计划推动机器人领域的跨国科研协作，并于 2024 年通过《人工智能法案》，构建全球首个全面的人工智能监管体系。2025 年，我国首次将“具身智能”写入政府工作报告，将其列为未来产业重点发展方向，体现出国家层面的高度重视。各地方政府也积极跟进，

通过建设人工智能产业园、创业孵化基地与数据采集中心等方式，为企业提供多方面支持。

全球资本市场的活跃也反映出该领域的前沿性。全球科技巨头纷纷进军具身智能领域，谷歌、微软、英伟达、特斯拉通过产业资源整合和战略投资纷纷入局，多家具身智能公司估值突破百亿元。风险投资机构对具身智能领域的投资规模呈现快速增长趋势，据披露数据统计，截至 2025 年 9 月，国内具身智能领域投资事件数近 500 起，融资总额已超 300 亿元人民币，投资方向遍及机器人硬件制造、核心软件算法、开发工具平台与垂直领域解决方案，其中，人形机器人更是成为全球关注的焦点。

发布方	政策文件	主要内容
美国	《国家机器人计划 3.0》	提供 1400 万美元的资金支持，主要研究集成机器人系统
中国	《人形机器人创新发展指导意见》	人形机器人有望成为继计算机、智能手机、新能源汽车后的颠覆性产品，并按照谋划三年、展望五年的时间安排，对 2025 年和 2027 年的发展目标做了战略部署
欧盟	《欧洲地平线》	2021–2022 年为机器人相关项目提供总计 1.985 亿美元的资金支持
德国	《2025 高科技战略》	为机器人在内的研究每年提供 6900 万美元的资金支持，到 2026 年总预算为 3.45 亿美元
日本	《机器人新战略》	2022 年的投入超过 9.305 亿美元，包括下一代人工智能和机器人的核心集成技术
韩国	《第三版智能机器人发展计划》	推动机器人成为第四次工业革命的核心产业，为《2022 智能机器人行动计划》投资 1.722 亿美元

表 1 为推进具身智能发展发布的国际政策文件汇总

1.3 具身智能的探索与挑战

作为跨学科前沿领域，具身智能整体处于发展初期，以 AI 大模型和传统控制理论为代表的两大学科正在碰撞融合，行业发展的核心问题逐渐聚焦在具身大模型——如何在动作层实现一个通用控制器，并因此呈现出多元探索路径与深层结构性挑战并存的复杂格局。

技术路线与能力正在探索中。语言大模型发展为具身智能提供了良好的任务理解和规划能力，但在动作控制层级，具身大模型的路线和结构仍在多元探索。当前，行业主要采用端到端的 VLA（视觉 – 语言 – 动作模型）模型路线，模型能力正在优化提升中。除了模型算法的选择，行业也逐渐出现系统融合、硬件能力不足等工程化问题，比如机器人“大脑”和“小脑”的分离导致协同延迟，而融合方案在算力分配与实时性上面临工程挑战；硬件本体主要因电机发热、磨损产生性能差异问题，以及现有感知、驱动和灵巧度均存在差距，仍需要大量科研和产业力量投入。

数据瓶颈问题突出。不同于互联网规模的文本与图像数据，具身大模型需要依赖收集特定机器人的操作数据，包括实际机器人本体的执行记录、模拟器生成的数据等，并且需要标注标签，数据采集成本呈指数级上升。同时，具身智能系统的高度复杂性导致算法与硬件紧密耦合，异构数据难以互通，已成为具身智能发展的重要瓶颈。特别是在灵巧操作领域，现有末端执行器的灵活性、触觉感知、稳定性等综合能力远逊于人手，成为实现精细物理交互的技术瓶颈。行业普遍认为要实现具身智能的涌现至少需要百万小时来自真实世界的物理互动数据，目前积累的数量仅不到 5%。现阶段实际可用数据量远未满足需求，且数据采集和使用方法尚未形成共识。此外，领域内缺乏统一的能力评估基准，使得不同方法的性能对比缺乏科学依据，严重阻碍了研究进展的可度量性。

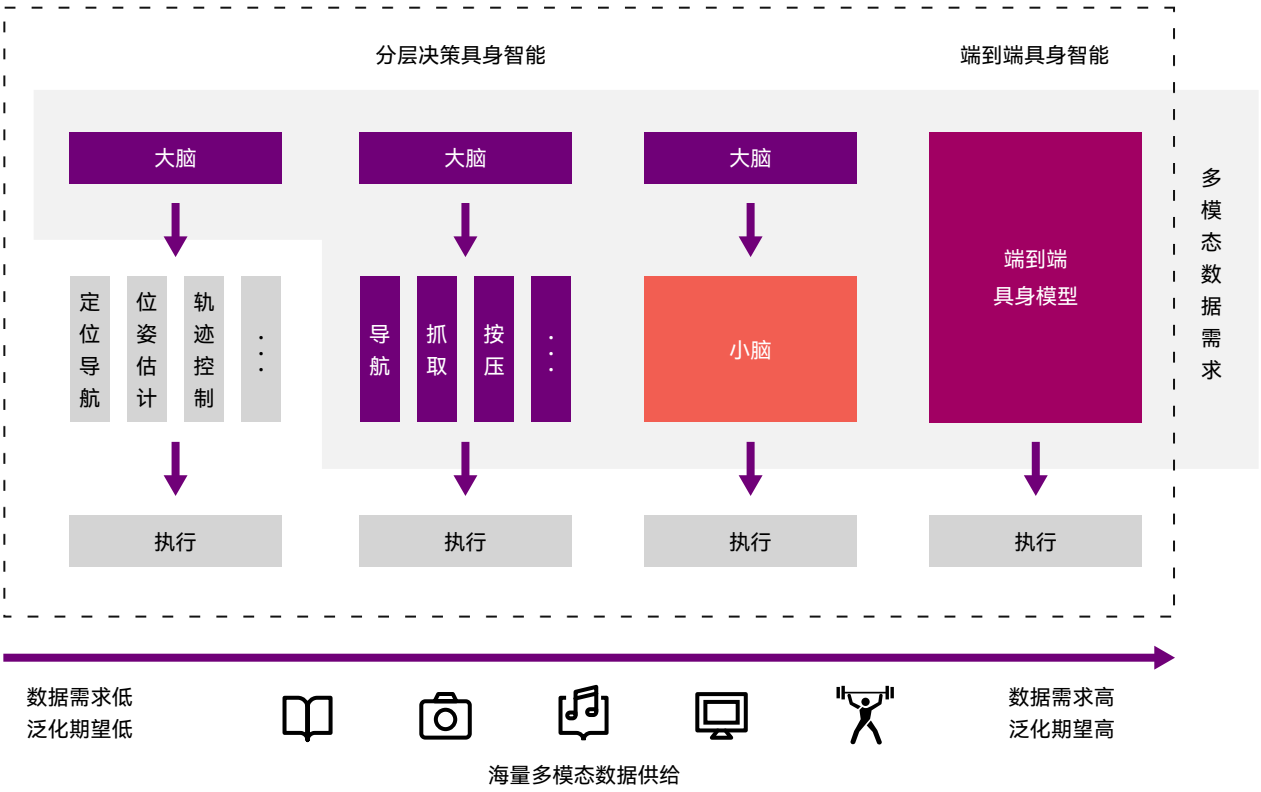


图 1 具身智能的技术架构与数据需求

成本与商业化仍需要时间。具身智能行业长期潜力巨大，但是短期面临成本高、场景窄、回报周期不明朗的现实压力。首先是硬件成本，核心零部件价格昂贵，且供应链尚未成熟，如六维力传感器、谐波减速器、电机等，产业生态的不成熟直接导致各环节缺乏统一标准，定制化程度高。其次，当前机器人在动态环境下的操作精度、响应速度和可靠性，距离大规模商用仍有差距。同时，商业模式单一，产业收入主要依赖硬件销售，服务等增值收入占比低，行业商业化创新略逊于技术发展。

发展具身智能是多领域融合的系统性工程，数据是跨领域的真实枢纽，贯穿全部链条。从数据视角出发，不仅可以观察不同技术路线在数据需求与处理能力上呈现的收敛或分化信号，追踪数据成本变化、甄别场景价值，更为投资与商业决策提供了至关重要的现实锚点出发。因此，本文希望通过梳理数据采集路线、研究数据使用方式，探究行业最底层的驱动力和最现实的制约因素。

02 具身智能的数据采集路线

具身智能技术正处于发展早期阶段，但是 VLA 模型首次大规模验证了“多模态大模型直接驱动物理动作”这一革命性思路的可行性，在未来某个时刻，具身智能算法的创新可能也会迎来类似“ChatGPT 时刻”的突破进展，在迈向更强大、更可靠的模型探索过程中，高质量、大规模、多模态的数据集的积累，是一条必须逐步填平的鸿沟。

当前，行业聚焦于如何进一步提升模型能力，获取数据成为首要问题。每种数据采集方式在数据精度、规模化以及多样性方面各有千秋，比如，遥操作数据能够获取高精度的多维度数据，但硬件成本和采集规模、速度上略显劣势；动捕数据兼具真实数据和合成数据的优点，正在补充操作和全身运动控制任务的数据需求；互联网视频数据和合成数据拥有巨大潜力，科研领域正在持续攻克应用中的难题。

本章节主要收集并归纳了部分重要科研原型和市场产品，并尝试讨论与其相关的数据使用场景和方法。

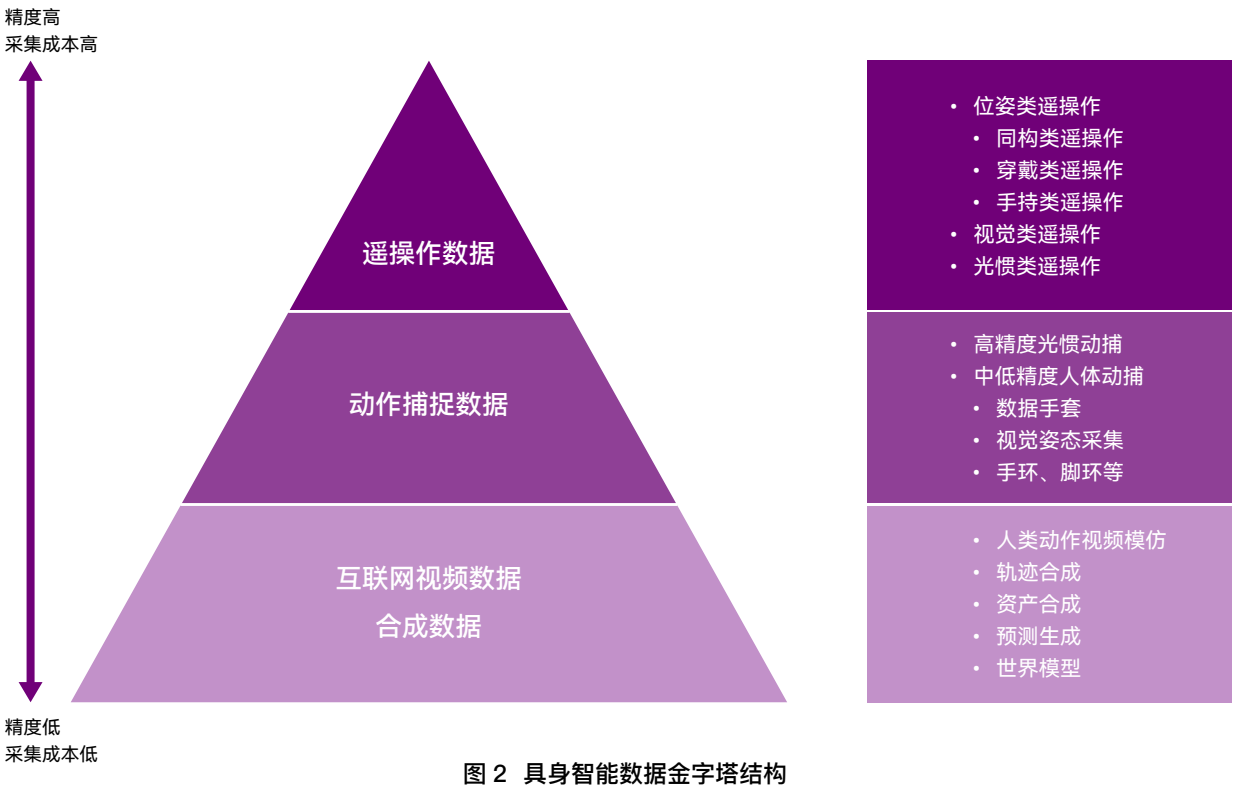


图 2 具身智能数据金字塔结构

2.1 遥操作数据

遥操作类（Teleoperation）数据采集在当前发展阶段内占据主流地位。由于机器人本体发展还不成熟，在现实环境中让机器人自己收集数据存在非常大的障碍，而使用与实际应用的机器人完全相同的机器人进行数据收集，可以避免物理形态的差异，并实现高精度。遥操作提供了一种最直接有效的人机协同控制方式，同时也是最可靠的高质量交互数据

来源，保证了数据的物理真实性——同步、完整地记录下整个操作过程中的所有物理状态变化，从而形成一条包含“动作意图 – 环境感知 – 物理执行”全链条的高保真数据轨迹。

因此，遥操作成为操作演示数据的主要来源，并且遥操作发展起步较早，形成了当前多种技术集成、不同类别的采集方案。

2.1.1 位姿类遥操作

位姿类遥操作特点是人类操作员通过直接记录位姿数据进行远程控制，其中，遥操设备需要将位姿信号转换为机器人的位姿控制信号。因此在遥操作设备中，位姿类遥操作设备最为丰富，可以划分为同构类遥操作、穿戴类遥操作、手持类遥操作三大类。

① 同构类遥操作

同构类遥操作是在两个完全相同的机器人之间实时动作复现，主从两台机器人的动力学结构完全相同，遥操作的控制和动作复现难度大大降低，并且允许操作员直观地控制机器人动作。

2023 年初，斯坦福大学的开源机器人项目 ALOHA（A Low-cost Open-source Hardware System for Bimanual Teleoperation）通过关节复制（Joint-copy）开创了低成本、高性能双臂遥操作范式，由于需要额外的机械臂作为用户的控制器，因此成本仍然较高。

不久之后，加州大学伯克利分校提出的 GELLO（A general, low-cost, and intuitive teleoperation framework for robot manipulators）遥操作系统，利用 3D 打印部件实现一个通用、低成本、直观的操作臂系统框架，成为机械臂的缩小版，进而推动遥操作技术发展。

2023 年末，斯坦福大学的研究团队继续推出 Mobile ALOHA，该系统降低成本的同时支持全身遥操作，能够胜任复杂的移动操作任务，使用四个摄像头可以覆盖较大的工作空间，并且操作员在整套设备的后方相对直观，结合高效的模仿学习算法，只需提供大约 50 次任务演示，就可以让机器人学习并完成各种任务。

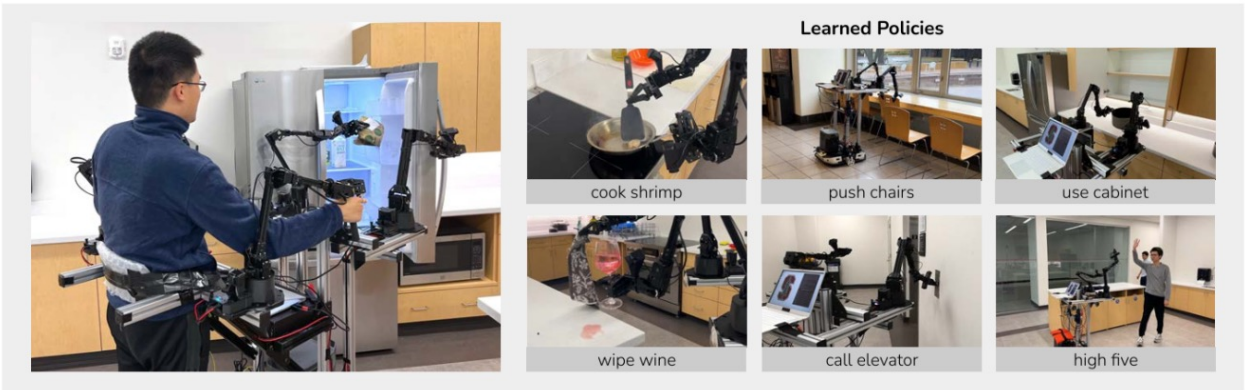


图 3 Mobile ALOHA 方案示意及操作演示

同构类遥操作范式主要采用关节复制（Joint-copy）方法，具有实时性好、延迟低等优点，尤其适合需要精确手动控制的任务。基于这种特点，松灵机器人推出 Cobot Magic 轻量级标准化双臂遥操作硬件平台，实现多摄像头与多机械臂的高帧率数据同步与采集，可采集标准 ALOHA、ARIO 等数据格式，同时适配 ALOHA 模型的训练与推理。

智元机器人亦在同构遥操作路径上进行了规模化实践。2024 年 9 月，其在上海张江启用了行业首个数据采集工厂，占地 4000 平方米，分割为家居、餐饮、工业等不同主题场景，还原真实世界的物件布局。厂区内每日有超过 100 台机器人同步训练，数据采集员通过同构遥操系统或 VR 设备精准控制机器人完成倒水、叠衣、分拣等任务，单机单日可产生上万条高质量轨迹数据。

② 穿戴类遥操作

相较于同构类遥操作，穿戴类遥操作提供了更高的直观性和自然性，通常用于需要精确控制和高度灵活性的任务。

2023 年 9 月上海交通大学卢策吾教授团队和上海人工智能实验室共同提出了 AirExo 可穿戴遥操作设备，核心目的是实现自然状态下的野外演示（In–the–wild）的数据采集，帮助机器人通过模仿人类的全身操作来学习复杂任务，但未能完全摆脱真机数据，由于原始数据存在不精确性，且依赖特定任务的真机数据微调，泛化能力也受到限制。

2025 年 3 月，卢策吾教授团队推出 AirExo–2 系统，旨在完全摆脱对实体机器人和真机遥操作数据的依赖，通过算法自动将不精确的人类动作数据转化为目标机器人可执行的“伪机器人演示”实现动作迁移，配合全新的策略网络（RISE–2），首次实现仅凭穿戴设备采集的人类演示即可训练出高性能策略，无需真机遥操作数据。

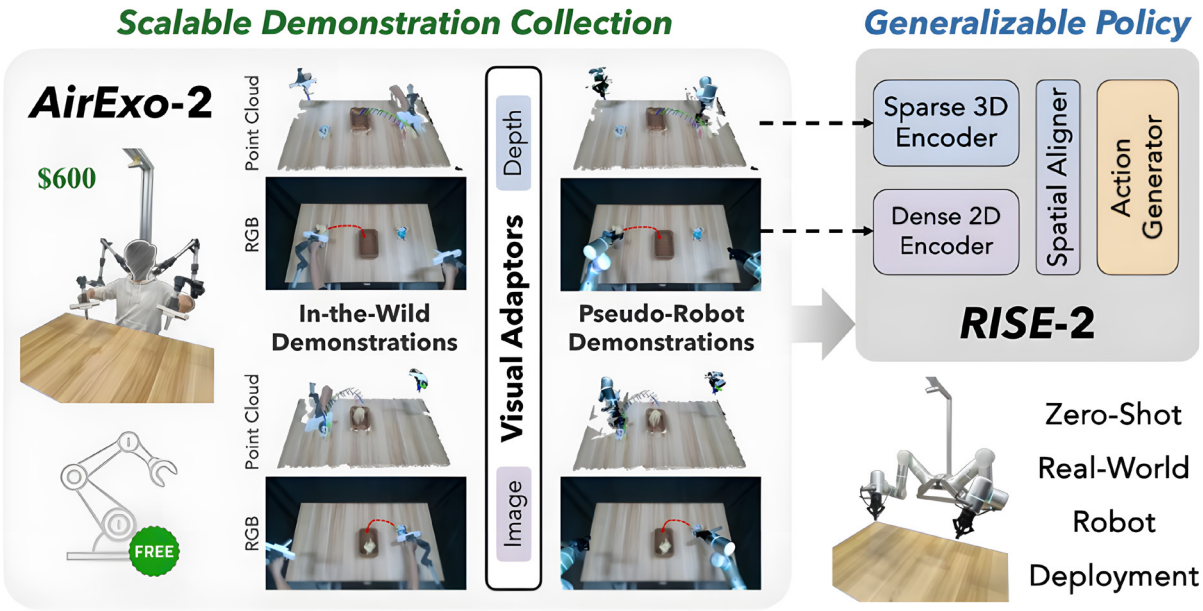


图 4　AirExo–2 系统与 RISE–2 策略网络结构

穿戴类设备与野外演示数据开辟了新的数据采集路径，逐渐延伸与视觉、动作捕捉等技术融合，为不同任务需求提供数据采集方案。比如，2024 年加州大学圣地亚哥分校提出开源遥操作设备 ACE（A Cross–Platform Visual–Exoskeletons System for Low–Cost Dexterous Teleoperation），是一种跨平台视觉 – 外骨骼系统，兼容各种机器人硬件，包括各种末端执行器，如夹持器和多指手，提供了灵活性。其外骨骼臂配备高分辨率编码器，可精确读取关节位置，确保准确的末端执行器跟踪。

从产业动态看，包含外骨骼在内的穿戴类遥操作设计，因其保留数据真实性基础上，场景适配更多、兼容性强以及成本可控，能更直接地解决具身智能商业化的核心痛点，从而成为更受企业和市场青睐的技术路径。

戴盟推出的 DM–EXton 系列便携穿戴式遥操作数据采集系统，采用无线连接，支持 54 个自由度动作捕捉，覆盖手

臂及手部运动范围，采样频率达 800Hz 以上，可适配主流机器人平台，核心优势在于轻量化设计（2.5kg）和高兼容性，能为机器人远程操控和智能训练提供高质量多模态数据支持。

灵巧智能推出 DexCap 系列数据采集系统，采用异构外骨骼架构，能实现毫米级动作捕捉与千赫级响应。

灵心巧手推出的 Open TeleDex 模块化机器人遥操作系统，主打“三重任意”（任意外接设备、任意机械臂、任意灵巧手）开放式架构，同时无缝支持接入 VR 设备、可穿戴式外骨骼等其他控制终端进行主端输入，推动解决设备兼容性差的行业痛点。

因时机器人推出的外骨骼力控手套集成了 5 个直线伺服驱动器，每根手指可提供 3N 的主动力和 5N 被动力，支持 14 自由度动作捕捉，价格区间为 3000–5000 元，其核心优势在于直驱技术带来的高精度和稳定性，已在工业领域实现大规模部署。

③ 手持类遥操作

手持类遥操作设备相对结构设计简单，获取位姿数据和转换得到的控制信号稳定度高，便于集成到系统，但很多情况下不能直观表达操作意图，更适合简单环境下的数据采集工作。

UMI（Universal Manipulation Interface）作为典型的手持类数据采集硬件工作，体现了“无本体数据采集”的思想，即目标机器人能执行相应的末端轨迹，就可以复用这些训练数据。

初代 UMI 确立了“手持夹爪 +GoPro 手腕摄像头”的核心范式，通过统一人类与机器人的“观察视角”，极大地减少了数据迁移的难度，让同一套数据能用于训练不同机械臂。这一范式验证了人类手部和世界交互的高保真、低摩擦数据的收集接口，也因此逐渐演化出了 Fast–UMI、MV–UMI、DexUMI、ActiveUMI 一整个家族，为手部操作提供了全面的数据接口支撑。

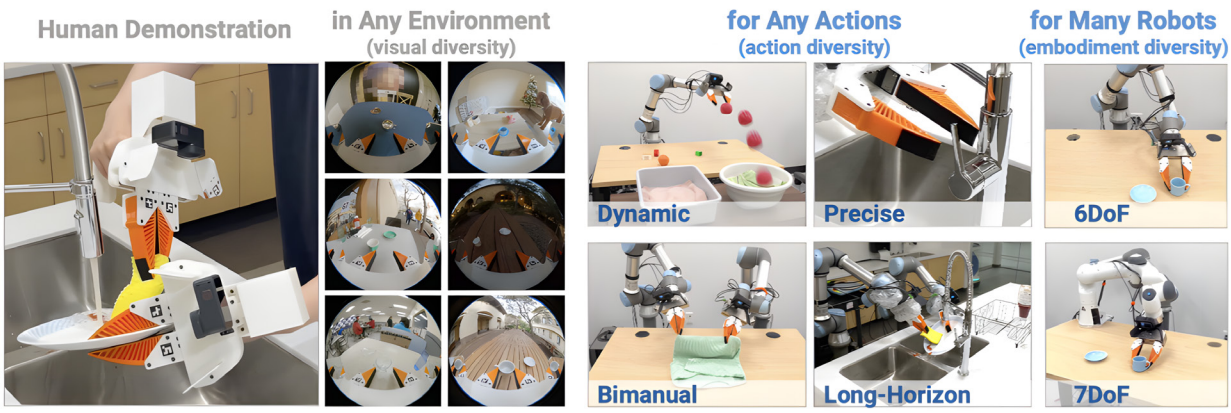


图 5 UMI 方案展示多种任务演示

手持类遥操作主要依赖末端位姿控制（End–Effector Control）技术，通过逆运动学（IK）直接控制机器人的末端执行器（如抓手或工具）的姿态和位置。操作者只需要指定末端的目标位置，不需直接控制每个关节，系统会自动计算关节角度来达到该位置。由于简化了操作者的工作，不必管理多自由度机器人的每个关节，因此也成为产业内的重要遥操作方式。

松灵机器人继 Cobot Magic 后，推出 PIKA 轻量化小尺寸手持夹爪，重量仅 550g，位姿精度最高达 1.5mm。除此

之外，由于松灵自研的 PiPER 机械臂具有较高通用性、较低开发难度和较低的价格，受到了开发者的欢迎，也演化出手柄、手环等第三方遥操作方案。

当前，鹿明机器人推出 FastUMI Pro 数据采集系统，将单条数据采集时间从 50 秒缩短至 10 秒，效率提升 5 倍，同时将综合成本降至传统方法的五分之一。同时，鹿明也在持续打造以人为核心、可灵活部署的“数据生产车间”，持续推动无本体 UMI 数据采集的落地。

2.1.2 视觉类遥操作

视觉类遥操作主要使用视觉传感技术捕捉人类操作员的动作，然后将这些动作转换为控制命令来操作机器人。由于利用了运动重定向（Retargeting）方法将人类动作映射到机器人动作上，动作映射的准确性依赖于重定向算法的质量，尤其在复杂场景下，动作精度和异构结构上需要较多调试。

NVIDA 和 CMU 开发的 DexPilot 系统，依赖 4 台 Intel RealSense 深度相机和 2 块 GPU，为高自由度（23-DoA）灵巧手系统提供低成本、高保真的遥操作方案，技术上主要使用了 DART（Dense Articulated Real-time Tracking）匹配手部模型与输入点云，实现实时跟踪操作员手部姿态和关节角度，验证了纯视觉方案可控制复杂灵巧手完成精细任务，证明了高质量视觉数据的可行性。

DexPilot 一定程度上是基于定制的工作空间，场景拓展有限，AnyTeleop 在机器人模型、部署环境普适性上进一步提升。AnyTeleop 开发了一个统一、通用的遥操作系统，能支持多种机器人手臂和灵巧手模型、多种现实环境、多种相机配置和多操作员协作。另外，AnyTeleop 在真实世界数据采集外，还可以用于仿真环境下的数据生成（虚拟遥操作），因此具有较好的通用性。

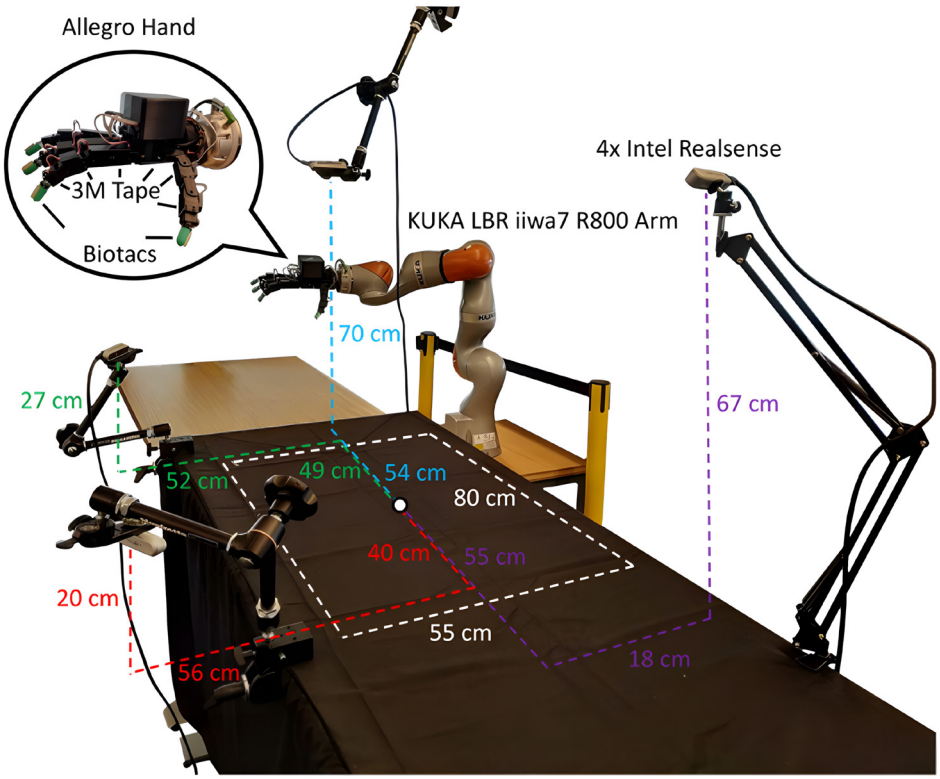


图 6 DexPilot 系统工作空间布局

视觉技术应用还存在一条极简路径，即单目视觉实现手部动作跟踪和机器人模仿，目前有许多研究工作正在挑战。

纽约大学开发的 DIME（Dexterous Imitation Made Easy），是一种基于学习的框架，可以高效实现灵巧操作。通过模仿学习，利用单目 RGB 相机采集人手动作，并将其映射到机器人手上，从而实现复杂灵巧操作任务的高效学习。

CMU 的研究 H2O（Learning Human-to-Humanoid Real-Time Whole-Body Teleoperation）提出了一种基于 RL 的全身遥操作框架，尝试了从人类运动数据（AMASS）到机器人控制的仿真数据预处理流程，通过用单个 RGB 摄像头实现实时全身遥操作，但存在精度不足问题。不久后，团队又推出了 OmniH2O（Omni Human-to-Humanoid），作为 H2O 的全面升级，解决了依赖动作捕捉获取全局速度的问题，能在室内外完成高精度、灵巧的全身移动操作任务，发布了首个真实世界人形机器人全身移动操作数据集 OmniH2O-6，并且可通过模仿学习从遥操作数据中训练自主策略。

同期，斯坦福大学研究团队发布 HumanPlus，旨在使用合成数据和遥操作数据，解决机器人全身运动控制问题。利用单目 RGB 相机实现“影子学习”，通过实时估计人体和手部姿态，并将人体姿态重新定向为人形机器人的目标姿态，输出是 19 维的仿人身体关节位置设定点，这些设定点随后转换为力矩信号。HumanPlus 的工作不仅证明了仅凭视觉就能实现对人形机器人高质量、全身的遥操作与数据采集，其两阶段的训练方式也为利用海量人类数据训练通用人形机器人指明了一条极具潜力的技术路径。

2.1.3 光惯类遥操作

光惯类遥操作是一种结合光学运动捕捉系统和惯性测量单元（IMU）的复杂操作系统，综合了穿戴类遥操作技术和视觉类遥操作技术的优势，可以实现对人类操作员动作的准确、连续、可靠跟踪。由于价格高昂，光惯类遥操作早期仅在影视、游戏制作等领域使用，近年来逐渐拓展到具身智能的数据采集领域。

光学和惯性传感设备已经发展多年，市场上有较多 VR、AR、Xsens 等产品，目前科研领域主要使用此类设备搭建遥操作系统。

由香港大学和加州大学联合开发的 Bunny-VisionPro 是一套实时双臂灵巧操作的遥操作系统，通过 VR 头显和低成本触觉反馈设备（ERM 振动马达），实现对高自由度双臂机器人的精确控制。同时，采集到的数据可以进行时间戳对齐和降采样，用于后续的模仿学习。

斯坦福大学开发的 DexHub 和 DART（Dexterous Augmented Reality Teleoperation）主要通过 AR 和云托管的仿真环境，实现大规模的机器人数据采集，并推动机器人学习的互联网化。使用 Apple Vision Pro 等 AR 设备，通过 RealityKit 将仿真环境中的机器人和场景以逼真的 AR 对象形式叠加到操作员的真实环境中，操作员实时操作机器人完成任务并采集数据，采集的数据自动存储在 DexHub 数据库中，以便随时下载或共享。

TWIST（Teleoperated Whole-Body Imitation System）是斯坦福大学和西蒙弗雷泽大学的研究团队在 2025 年提出的一项创新技术，旨在解决人形机器人全身协调控制的难题，该系统使用动作捕捉设备精确记录操作者的全身运动数据。然后，通过算法将人体运动学模型映射（Retargeting）到目标人形机器人的关节空间，生成可供机器人跟踪的参考动作片段，更为关键的创新是 TWIST 结合了强化学习（RL） 和行为克隆（BC），训练出一个单一、统一的神经网络控制器，不仅能高精度跟踪重定向后的动作，还能自主保证机器人在执行时的平衡性与稳定性。

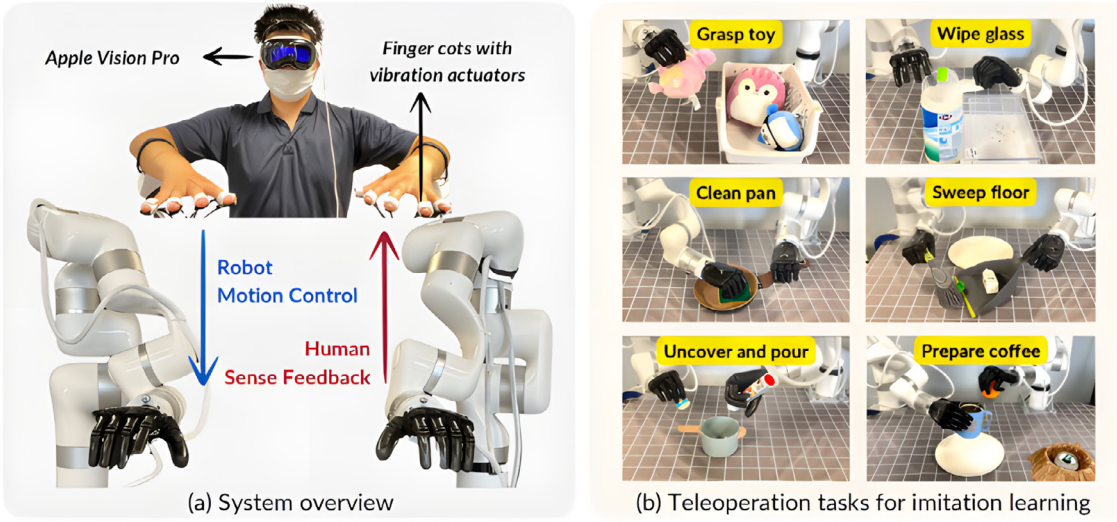


图 7 Bunny-VisionPro 系统示意

利用 VR 设备的虚拟现实控制（VR-based Control）为操作者提供沉浸式体验，易于理解和操作，提高控制的灵活性和精度，也更加适合非专业操作者。因此许多业内企业采用这种方案，如 Optimus 的早期方案，通过 VR 遥操和动捕手套（如 Xsens、Manus）实现人机动作映射，数据质量最高但成本也最为昂贵，仅整套设备成本就在 20 万元以上。国内头部具身智能企业智元机器人也采用类似方案，使用 VR+ 全身动捕设备实现遥操作，全程人工在环，采集并开源了百万真机数据集 AgiBot World。

另一方面，由于影视游戏需求，产业内一直存在高精度动捕设备及数据提供商，具身智能的数据需求为动捕数据服务企业带来新的商业场景。

诺亦腾（Noitom）创立于 2012 年，专注于人体动作数字化技术的研发与应用，构建了从开发平台到垂直应用的一揽子解决方案能力，Perception Neuron 系列产品已广泛应用在多个领域。

青瞳视觉自 2015 年起，自主研发并生产具有国内外领先水平的红外光学动作捕捉系统，从影视动画制作领域，逐渐开始服务于机器人领域的研究及开发工作。

艾欧智能作为具身智能数据服务企业，也推出了全身动作捕捉解决方案，包含高分辨率摄像头与深度传感器相结合的头盔，全身配备 14 个 IMU 传感器，手套 22 个 IMU 传感器，提供全身动作与环境的信息采集。

图 8 诺亦腾手指惯性动捕方案

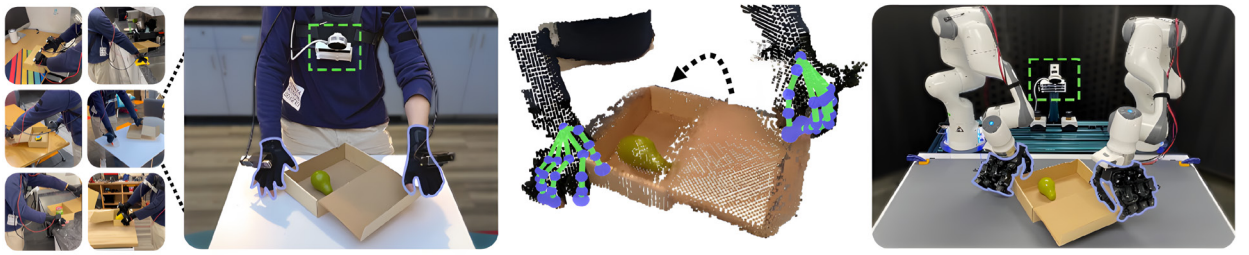
2.2 动作捕捉数据

基于前文对遥操作数据方案的梳理，视觉类遥操作、光惯类遥操作均使用了动作捕捉相关技术，当前越来越多的研究和产业工作正在基于人类动作搭建数据采集系统，而获取的动作捕捉数据既非纯粹的物理交互实录，也非完全虚拟的参数合成，它占据了一个独特的战略中间位置。所以，本文将动作捕捉数据单独进行分类，尝试探究如何使用这种连接真实物理世界与数字仿真世界的关键桥梁。

动作捕捉（Motion Capture）指的是将人的动作和姿态数字化的过程，这个定义本身与使用的方式和方法无关，无论是通过计算机视觉、传感器、磁追踪，还是外骨骼等不同技术手段，只要能把人的动作姿态捕获下来、转化为数字序列，都属于动作捕捉。

对于具身智能领域，动作捕捉数据可以划分为人体动捕示教数据和人类视频演示数据，获取方式主要有三种方法：基于相机的视频或图像、基于 VR、基于动捕设备，这三种方法获取数据的质量和成本依次递增，但数据规模依次递减，所以实际使用中，根据具体需求会在模型不同的训练阶段引入。

相比于遥操作，动捕数据示教侧重于使用各种动捕设备（如数据手套、动作捕捉服等）记录操作员的动作，并为机器人作示教，而不是精确对应机器人运动，在采集的数据可用性上存在差异。



(a) DexCap: Portable motion capture system　(b) Mocap data and 3D scene　(c) DexIL: Dexterous imitation learning

图 9　DexCap 系统方案示意

斯坦福大学李飞飞团队于 2024 年发布的 DexCap 系统，是一套旨在解决机器人灵巧操作数据采集难题的便携式手部动作捕捉系统。DexCap 包含一个可穿戴的相机背心、两个手套上的 SLAM 相机和一个胸前的 RGB–D LiDAR 相机，核心在于低成本、抗遮挡地采集高质量人类手部动作数据，能够在各种真实环境中快速采集高质量的 3D 手部运动数据，支持复杂任务，并直接用于训练机器人模仿学习的策略。TWIST2 采用 VR+ 手柄 + 腿环来替代光学动捕设备采集人体动作，操作员单人即可完成全身操控，在低成本、高效率采集数据的基础上，实现了学术界首次基于视觉的全身数据采集系统。同时，TWIST2 提出了一个分层视觉运动策略框架，能够让机器人能摆脱遥控，完全自主执行任务。

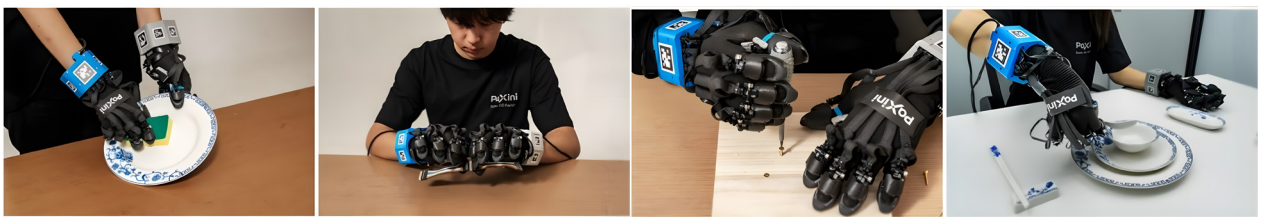


图 10　帕西尼感知 PMEC 数采方案

帕西尼感知（PaXini）基于自身在多维度阵列式触觉感知领域的技术积累和量产经验，自主研发了数据采集设备 PMEC，包含约 10 个触觉传感器，实现操作员手部的多维力采集。同时，帕西尼已在天津落地了 12000 平方米的具身智能超级数据工厂（Super EID Factory），其内部部署了 150 个标准化采集单元，每天最多可采集 55 万条数据，预计每年生产近 2 亿条融合触觉、视觉、文本等多模态的高维数据。

它石智航主张利用 Human-Centric 数据，通过硬件层面的创新，构建了一套轻便、模态齐全、可穿戴的具身数据采集系统 SenseHub，同时在灵巧手操作终端拥有 TARS Dex 灵巧手和 TARS Dex 视触觉夹爪，使算法能力能够更真实稳定地映射到物理世界。

诺亦腾（Noitom）Perception Neuron 动捕系统通过生成数字人运动数据，为控制模型学习提供示教数据。相比传统的光学动捕系统，PN 系列成本更低，易于普及，以其高精度、低延迟、抗磁干扰等特点，在具身智能数据采集领域受到各类研发机构的关注。该系列动捕产品包含不同级别，可以满足不同用户需求：入门级惯性动捕系统 PN3 包含全身及手部 27 个传感器，采用超小型无线惯性传感器，重量仅为 4 克，适合动画游戏开发；专业级无线惯性动捕系统 PN Studio 使用了超高精度的航天级传感器标定方式，能精准追踪手指的精细姿态，支持在 1000 平方米范围内实现最多 5 人全身和手指的动作捕捉，以及 20 个道具的实时追踪；PN Hybrid 提供基于光学与惯性多源融合核心技术打造的创新型动捕平台。2025 年诺亦腾机器人跨本体数据工厂在深圳市龙华区正式揭牌运营，进一步推进无本体数据采集的规模化，让数据采集与机器人本体解耦，直接将全量传感器（视觉、力触觉、深度视觉等）穿戴在操作者身上，以超高精度捕捉人类在真实操作中的全模态数据。

PN S 惯性传感器	17 个
PN S 备用传感器	1 个
PN S 动捕手套	1 双
PN S 传感器充电盒	1 个
PN S 数据收发器	1 个
充电盒电源线	1 套
PN S 全身绑带	1 套
Type-C 充电线	2 条
防爆手提箱	1 个
全向天线	1 条
Axis Studio 软件	1 套

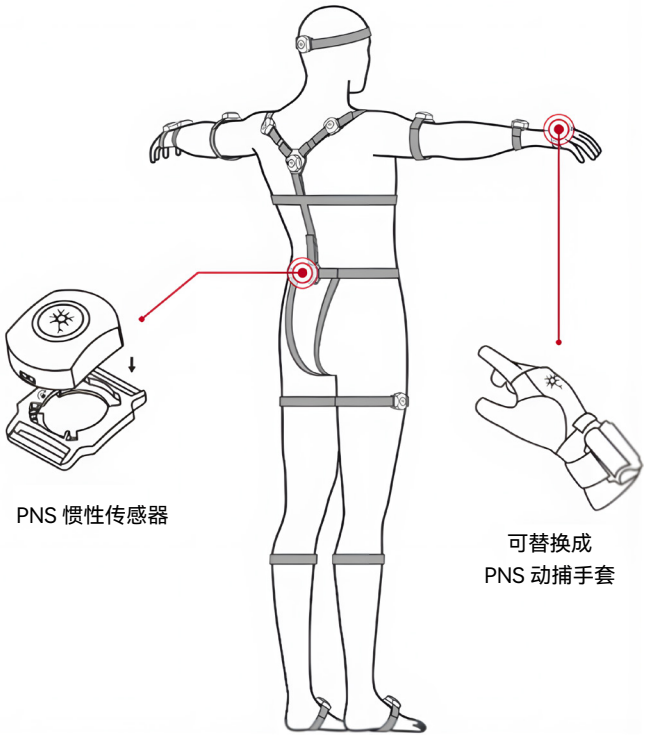


图 11 诺亦腾 PN Studio 方案构成

2.3 互联网视频数据和合成数据

2.3.1 人类视频演示数据

与文本数据类似，互联网也拥有海量的视频数据，其中包含人类动作的视频也具有演示作用。人类视频演示学习（Learning by Human Video Demonstrations）是一种新兴的机器人学习方法，核心是利用人类视频演示作为知识来源，使机器人理解并模仿人类行为来完成复杂任务，无需手动编程或大量的机器人数据采集。因为无需大规模的机器人数据采集，该技术相比使用机器人进行专家演示更经济，但对算法设计和训练的要求更高，目前提取的动作仍不精准，无法直接跨本体使用。

在该领域，当前较为前沿工作有 MimicPlay、HumanPlus、ATM、VideoMimic、GR-2 等等。

MimicPlay 是斯坦福大学李飞飞团队和 NVIDIA 提出的基于人类视频的模仿学习框架，设计了一个分层框架，高层规划模型从人类视频中提炼“做什么”的抽象计划，底层控制策略模型则用少量精准的机器人示教数据学习“如何做”，配合 Dexcap 可以进一步提升动作精度。

HumanPlus 的创新在于结合真实数据与合成数据，利用大量人类视频数据将人类姿态（SMPL-X）映射为机器人可执行的关节角度，HST 实时遥操作生成高质量的“状态 – 动作”演示数据对，然后进行离线模仿学习训练 HIT 模型，仅需约 40 次演示即可形成自主策略。ATM（Any-point Trajectory Modeling for Policy Learning）提出“任意点轨迹建模”框架，从视频中预测物体或身体部位在未来任意时间点的轨迹。这种密集的轨迹预测作为一种强大的中间表示，能显著提升下游机器人策略学习的样本效率和泛化能力。

ATM 为“观看视频学习技能”提供了一个强大的预训练或特征提取工具，但是预训练模型无法直接应用，必须经过额外的数据采集进行微调（fine-tuning），所以在操作类任务相对更适合，配合 UMI 等方式采集底层控制数据，能够实现良好的控制策略。

VideoMimic 让机器人仅通过观察随手拍摄的单目人类视频，学习与环境适配的全身技能，其关键是从单目视频中联合恢复人体运动与场景的三维几何信息。VideoMimic 工作可以分为两大部分，首先是人与环境的三维重建，其次是人体映射数据的训练，VideoMimic 先基于少量、高质量的人体动捕数据进行预训练，训练出一个具备初始运动能力的 base policy，然后再利用第一部分的三维重建数据进行专项强化学习，每条动作序列都训练出一个对应的教师模型（teacher policy），最终将这些具体动作的教师模型蒸馏（distillation）为一个统一的学生模型（ student policy）。

字节跳动公司推出的 GR-2（Generative Robot 2.0）是一种创新的机器人大模型，采用了类似 LLM 训练的预训练和微调过程。在预训练阶段 GR-2 观看了 3800 万个互联网视频片段，实现视频生成和预测，微调阶段使用了 4 万条机器人轨迹数据，进一步提升动作预测能力，最终在超过 100 个任务中实现了平均成功率 97.7%。

Figure AI 在 2025 年 9 月推出的“Project Go-Big”计划，其核心方案是通过构建互联网规模的“机器人行为 YouTube”数据集，并训练一个能直接从人类视频中学习机器人控制策略的统一模型（Helix），来拓展机器人的语义覆盖能力。特斯拉也开始探索视频学习方案，通过录制员工执行任务的视频，提取动作信息训练 Optimus 机器人。

国内企业枢途科技也发布了自研 SynaData 数据管线算法，成本将至行业平均水平的千分之五，成本相较于遥操作降低 200 倍。目前，枢途与清华大学联合推出大规模多模态数据集 HORA，初期包含 15 万条轨迹，是全球首个基于视频转化的 3D 具身开源数据集。实测表明，仅用 10 条 HORA 数据微调模型，其效果可与 10 条真机遥操作数据相当；而使用 5000 条数据进行预训练后，模型在复杂任务上的成功率得到大幅提升。

近期，蚂蚁数科天玑实验室团队研发的 AoE（Always-On Egocentric）将“人 + 手机”转化为可持续运行的轻量数据节点，以极低成本实现高质量的人类视频演示采集。通过一台手机和一个低于 20 美元的颈挂式支架，实测中，针对宇树 G1 机器人的关电脑任务，仅靠 50 条遥操作数据时成功率为 45%，而引入 200 条 AoE 采集的真实数据后，成功率跃升至 95%。在数据匮乏时，AoE 承担了“启动学习”的关键补位角色。

2.3.2 合成数据

在具身智能领域，学术界和产业界对仿真数据、合成数据一直保持较高的关注度，生成这些数据的方式主要沿着两条技术路径展开：数据仿真（Data Simulation）与数据合成（Data Synthesis）。前者致力于构建高保真的动态物理环境，以孕育出符合真实世界交互规律的智能体行为策略；后者则侧重于以算法驱动，高效生成大规模、多样化的静态或序列化标注数据，以攻克感知与模仿模型的数据瓶颈。二者互为补充，共同构成了仿真数据基础设施的一体两面。

数据仿真是指通过计算机模拟技术生成虚拟环境和场景，核心在于创建一个虚拟世界，依赖高保真的物理引擎和场景模型，遵循物理规律，模拟真实世界中的物理过程和交互，生成用于训练机器人的数据。在实际使用中，数据仿真产生的大规模数据一般不会被全部储存，而是采用边生成边训练、训练完即丢弃的模式对模型进行训练，大多数机器人行走的在线强化学习使用了该方法。当前四足机器狗、轮式机器人等机器人，其全身运动控制能力大多基于强化学习算法，由于运动控制主要与环境刚性交互的特点，非常适合在仿真环境下进行训练。因此，一般的运动任务并不需要构建专门的数据集。而人形机器人由于自由度较高，对特殊任务要求需要补充专门的数据集训练，这些数据一般来自人体动作捕捉、人体姿态视频和合成数据，为机器人运动控制的学习提供了丰富资源。

数据仿真的优势与劣势都极为鲜明。仿真环境允许使用强化学习、进化算法等需要海量交互数据的搜索型方法，智能体可以通过并行化在成千上万个环境副本中同时探索，快速覆盖状态空间，发现人类难以设计的复杂或反直觉策略。与此同时，构建高保真仿真环境成本高，存在仿真到现实的差距（Sim2Real Gap）。由于仿真是对现实世界的简化建模，误差必然存在，这其中包括物理简化、感知简化、动作简化等。当前，在控制策略设计、动力学建模方面有相关优化方法，比如领域随机化（Domain randomization）、领域自适应（Domain adaptation）等方法可以增强策略学习的稳定性。但是，以电机等硬件为核心的执行器，存在响应延迟、饱和、齿隙和热效应等问题，且缺乏有效的传感机制，导致仍无法被充分模拟。

数据合成是指将算法、统计模型或真实世界数据导入仿真平台，然后通过算法合成在统计学上与真实数据相似的新数据的过程。数据合成弥补了真实数据的多样性，更加适合需要大量数据训练模型的场景。数据合成又可以进一步划分为轨迹合成、资产合成、决策生成和预测生成。

① 轨迹合成

轨迹合成核心服务于训练操作能力，目的是生成数据以训练得到相应的策略模型（Policy model）。轨迹合成主要包括路径规划和运动控制，即生成从初始位置到目标位置的平滑、连续、避障的路径，并且确保末端执行器按预定轨迹精确运动，满足速度、加速度和抖动等约束条件。

实践中，有两种轨迹合成方式。

一种方式是基于虚拟遥操作，依靠外部设备向仿真平台发送远程控制命令，以生成机器人的行为控制数据，但执行效率较低，偏向于遥操作数据采集，在大规模数据生成上存在短板。

另一种方式基于策略模型，使用已有的策略模型在仿真环境中自动合成大量数据可以显著提升效率，并可以构建强大的数据飞轮。

以 NVIDIA 推出的 MimicGen 方案为例，首先利用遥操作采集数据训练得到一个初始策略模型，然后将这个初始策略接入仿真环境生成大量轨迹合成数据，通过指标体系筛选出合成数据集，最后使用合成数据集继续训练初始策略模型，形成增量学习。

NVIDIA 公司后续推出的 DexMimicGen 轨迹合成方案，希望通过少量人类演示自动生成大量、多样化的数据集，门针对人形、双臂灵巧机器人，仅需 5 次人类演示，即可生成 1000 个双手灵巧任务演示，使用生成的数据训练机器人在多种任务中成功率显著提高。

亚马逊 FAR 团队首个人形机器人研究成果 OmniRetarget 也提供了一种人体动捕数据的轨迹合成方式，作为数据仿真的前置环节，将原始演示的数据转换成当前机器人可执行的运动学轨迹，解决了从人类到机器人的动作迁移与泛化，其生成的轨迹为下游的强化学习策略训练提供了高质量、物理合理的种子数据。

② 资产合成

资产合成主要是通过生成式 AI 和相关技术创建虚拟场景和物体，尤其是在仿真环境中可交互的对象，用于训练机器人的感知和交互能力。资产合成通常基于真实场景或物品，以避免生成脱离实际的任意资产，常用方法主要有数字孪生（Digital Twin）、数字表亲（Digital Cousin），涉及神经辐射场（NeRF）技术、高斯喷射（Gaussian Splatting）技术等 3D 点云处理或 3D 重建方法。

在具身智能领域，数字孪生常被用于合成与真实世界尽可能一致的可交换物体。

由哈尔滨工业大学深圳校区等单位开发的 RoboGSim 仿真平台，基于 Real2Sim2Real 范式，通过高保真的 3D 重建（3D 高斯喷射技术）和物理引擎，能够生成具有真实纹理和物理特性的合成数据，用于机器人策略学习和评估。

香港大学等研究机构开发的 RoboTwin 是一个创新的生成式数字孪生框架，通过结合 3D 生成时基础模型和大语言模型，为双臂机器人任务生成多样化的专家数据集。该框架能够从单个 2D 图像创建物体的数字孪生体，生成逼真可交互的场景，并通过空间关系感知的代码生成框架，结合物体注释和 LLM 分解任务，生成精确的机器人运动代码。

国内具身智能企业中，银河通用是使用仿真合成数据的代表，公司自研全仿真合成数据生产管线，在 NVIDIA Isaac 等平台基础上，通过程序化生成和物理渲染，创建海量的虚拟物体和场景，基于十亿量级的仿真数据发布了全球首个全仿真预训练具身大模型 GraspVLA，以及身智能灵巧手多样抓取仿真数据集 DexonomySim。

在合成数据领域内，光轮智能主张以“物理真实”为核心，构建 Real2Sim2Real 的合成数据闭环，使用物理设备采集真实物体参数真实物理参数，反向输入仿真模型进行建模，再生成数据训练机器人。目前，公司客户包括 Figure AI、智元机器人、银河通用机器人、DeepMind、字节跳动等，并与 NVIDIA 生态深度绑定，成为 NVIDIA GR00T 人形机器人基础模型的重要数据合作伙伴，并与其共同开发下一代仿真平台 Isaac Lab Arena。

群核科技通过酷家乐等工具，经过十余年积累了超过 5 亿个 3D 结构化场景和 4.4 亿商品模型的庞大资产，构成了群核空间智能平台（SpatialVerse）的基础。目前公司已与智元机器人、银河通用、穹彻智能、智平方等一批头部具身智能企业达成合作，为其提供仿真训练数据。

生境科技基于端到端空间 AI 生成的技术底座，以自研的“空间生成与理解的通用底座”为核心，生成大规模、高质量、带语义的 3D 空间资产，并将其同时应用于商业落地和为具身智能提供合成数据。公司从零开始搭建了独创的空间编码方法与自监督训练体系，以此作为生成和理解空间的基础，致力于打造一个通用的、平台级的 3D 空间合成数据引擎。

③ 决策生成

决策生成是具身智能分层决策技术路线中不可缺少的一环，主要利用大语言模型的理解和生成能力，实现从自然语

言指令到可执行动作指令的转换，一般包括任务分解和代码生成两个阶段。例如，卡耐基梅隆大学等机构提出的“LLM–Planner”框架，便让大语言模型扮演了“策略师”角色，针对“整理散落的玩具”这类指令，模型能自动将其分解为寻找玩具、抓取、放入收纳箱等子任务序列，并生成可直接在仿真或真实机器人上执行的 Python 代码，清晰地体现了从理解、规划到代码落地的完整链条。

决策生成技术被广泛应用于导航任务、复杂任务执行和人机协作，通常通过智能体（Agent）实现，将 LLM 接入仿真环境数据生成系统，生成相应的决策数据。其中，ViLa（视觉 – 语言 – 动作）模型架构是一种广泛采用的模型架构范式，通过将强大的视觉 – 语言基础模型与机器人策略网络相结合，使得智能体能够直接根据图像和自然语言指令生成动作，实现了高层次的任务理解与低层控制的深度融合。

同时，另一类工作如 CoPa（协作策略获取），则侧重于研究机器人如何通过模仿人类的协作演示，学习到可组合的协作技能，展示了决策生成技术在复杂人机、多机协作场景中的潜力。

④ 预测生成与世界模型

在具身智能的数据合成技术中，预测生成占据着独特而关键的位置。模型对于真实世界的理解能力通常难以衡量，将其转换为对事物发展的预测能力更为直观。预测生成的根本价值在于将机器人的学习从“状态 – 动作”的被动映射，升级为包含后果预估的主动决策。通过预测不同动作可能引发的未来状态，机器人可以在“想象”中评估和选择最优策略，从而处理需要长时序推理、规避风险或实现复杂目标的任务。

为了训练模型对事物发展的预测能力，一方面需要大量真实世界物理变化过程的数据，另一方面需要借助专业的生成工具生成真实世界难以采集的合成数据。而打造这种专业的生成工具，主要有两类方式：

一种是基于生成模型的预测生成，如生成对抗网络（GAN）、变分自编码器（VAE）等模型，逐渐被应用在生成丰富、真实的交互场景，侧重于视觉领域，尤其是视频的生成。谷歌等联合发布的 Gen2Act 方法，通过结合人类视频生成技术和机器人策略学习，能够显著减少对真实数据的依赖。MIT 等研究机构提出的 HMA 模型则从机器人运动视频生成角度，结合异构数据预训练和掩码自回归技术，能够处理来自不同机器人实体、领域和任务的多样化数据，并生成具有高保真和可控性的视频。

另一种是基于世界模型（World Model）的预测生成，世界模型概念源于人类内在的心智模型，通过感官获取抽象信息，在大脑中转化为对周围世界的具象理解，当前这个概念定义仍不清晰，一般认为世界模型的核心能力是对环境动态的预测建模。

世界模型需要结合强大的生成能力和对物理世界的理解能力，生成符合物理规律的内容。当前使用扩散模型（Diffusion Model）、Transformer 架构或贝叶斯网络等深度学习技术，能够生成高质量的视频、3D 场景等虚拟内容，例如，OpenAI 的 Sora 模型不仅能够生成丰富视频，还因其潜空间向量的使用具有一定物理性质；DeepMind 的 Genie 模型允许用户输入图片和操作指令生成一个可交互的虚拟环境；2024 年，李飞飞教授等创办的 World Labs 推出了全球首个空间智能模型，通过一张静态图片生成一个逼真的、可交互的 3D 世界。

NVIDIA Cosmos 也是该领域的重要模型，主要面向物理 AI 开发者，提供了一系列预训练的生成式世界基础模型，开发者可以直接使用他们生成合成数据，也可以使用 NVIDIA NeMo 微调，目的是加速自动驾驶汽车和机器人等物理 AI 系统的开发。

流形空间（Manifold AI）是国内探索世界模型技术路线的初创企业，独创 WMA（World Model Action）路线，区别于主流 VLM 范式，强调用世界模型作为机器人的基础模型，自研通用空间世界模型 WorldScape，具备“推理 – 想象 – 行动”三位一体能力，能够根据单张图片预测并模拟空间内的物理反馈。目前，流形空间已在无人机领域实现落地突破。

苏昊教授领导的 Hillbot 团队将学术上深厚的模拟器研发积累（如 SAPIEN）与前沿的生成式 AI 技术结合，利用 SAPIEN 等模拟器为生成的 3D 对象赋予物理属性，机器人可以与物体进行仿真交互，从而自动生成大量的“状态 – 动作 – 结果”数据轨迹，实现了从 3D 内容生成到物理交互仿真的全链条构建。

本章系统梳理了具身智能的三条核心数据获取路径：遥操作数据、动作捕捉数据与合成数据。当前，这三条路径并非孤立存在，已经开始出现融合演进的趋势，在从使用真实数据走向创造合成数据的过程中，伴随数据规模的增加和数据成本的降低，是行业对算法和数据处理能力的不断提升，是对工程化能力的新挑战。

03 自动驾驶的数据发展经验

自动驾驶历经十余年发展，可以看作是一种已实现规模化部署的轮式具身智能形态，展现出了从专用自动化向通用自主性发展的趋势，其实践验证了物理实体如何通过持续环境交互与数据驱动，这为通用具身智能提供了可资借鉴的宝贵经验。尤为重要的是，自动驾驶产业在迈向完全自主过程中所经历的技术选择与数据体系，历经了从严重依赖纯真机采集的静态数据，到仿真生成与真机验证结合的根本性转变，为具身智能的发展提供了重要参考。

3.1 高精地图：静态真实数据的经验与教训

高精地图是自动驾驶早期依赖的、典型的“静态真实数据”典范，其经验深刻揭示了纯真机采集模式的固有瓶颈。

在自动驾驶发展初期，车辆感知能力有限，无法满足高安全可靠性要求。早期感知算法在复杂光照、天气与路况下表现极不稳定，仅凭车载传感器难以实现厘米级精确定位与对环境要素的稳定感知。

在此背景下，高精地图作为一项关键技术被引入，迅速成为高级别自动驾驶系统的重要组成部分。高精地图巧妙地将“实时理解环境”这一复杂的视觉问题，转变为了“在已知地图定位”的相对简单问题，降低了对实时感知算法的依赖，使车企能够基于尚不成熟的感知系统快速搭建稳定的演示系统，加速了技术早期落地。

然而，高精地图的广泛应用也带来了技术路径依赖。

首先，制作覆盖广泛区域的高精地图需要专业采集车队与复杂后期处理，前期投入大。

然后，道路环境处于持续变化中，施工、改道与交规更新频繁，维持地图“鲜度”需要持续投入，对商业模式构成挑战。更为关键的是，依赖高精地图的自动驾驶车辆只能在已测绘区域运行，限制了其泛化能力，难以应对未测绘区域，与全域自动驾驶的长期目标存在矛盾。

此外，长期依赖地图提供的明确规划指令，可能削弱感知系统攻克复杂场景的动力，一旦脱离地图或地图出错，系统表现将显著下降。

高精地图从特定历史阶段提升效率、弥补短板的工具，因其显著效果而逐渐演变为系统的基础依赖，最终因规模化成本与泛化能力限制而被行业重新评估。

为破解困局，行业转向使用由众包车辆自动生成的轻量化地图，利用海量装备普通传感器的产线车辆在日常行驶中实时回传变化信息，经云端融合处理后，以低成本、高效率实现地图的动态更新。这一转变的本质是将数据采集从昂贵的、专门的生产活动，转变为嵌入到大规模日常应用中的“影子模式”，实时追踪人类驾驶员行为，形成模型训练与反馈的闭环。

高精地图的经验深刻警示，地图是静态的，场景是动态的，具身智能不能仅依赖实验室或工厂预采的固定数据集，真正的智能体现在对未知环境的适应。此外，构建一套类似“影子模式”的动态、闭环数据采集系统，将为具身智能破解规模化载体不足的困局提供了关键思路，尤其是人形机器人目前缺乏规模化部署载体，数据采集依赖高成本原型机与有限场景，导致发展陷入循环制约。

3.2 数据异构融合：分层采集与合成

自动驾驶的系统架构需求驱动了数据形态变革。从早期为各独立模块提供精确信号，到为融合感知模型提供特征地图，再到为端到端大模型构建动态世界表征，不同功能层之间的数据融合，随算法范式的升维而不断深化。

在模块化架构时代，数据采集的核心任务是保证感知、预测、规划等独立模块的稳定输入。激光雷达、摄像头、毫米波雷达等异构传感器需进行精密的时间同步与空间标定，实现数据级的前融合。部件厂商在这一底层工作中扮演了关键角色。禾赛科技在 2017 年与百度 Apollo 联合发布的 Pandora 一体化传感器中，开创性地将 40 线激光雷达与多个摄像头集成于单一圆柱体结构中。Pandora 通过一体化的设计，将激光雷达与摄像头的相对位置固定，解决了此前分立式传感器需要人工标定的难题；同时，禾赛科技负责把控激光雷达与摄像头的触发时机，确保两个传感器信息采集的同步，将传统方案中可能存在的数百毫秒误差压缩至微秒级。速腾聚创在 2018 年 CES 上推出的 LCDF（Lidar–Camera–Deep–Fusion）技术，将 MEMS 固态激光雷达与摄像头进行硬件上的底层融合，让自动驾驶车辆能全方位感知真实世界的三维空间色彩信息。其核心价值在于解决了多传感器数据的时空一致性难题——传统的做法需要下游厂商单独对分立传感器进行标定，费时费力且难以保证时间同步与空间校准。LCDF 技术让两者预先融合，保证了两者的时空一致性，使自动驾驶车辆在做决策算法之前，就能实时感知并处理相关信息。

随着深度学习兴起，感知模块率先模型化，感知层需要海量带标注的真实图像与点云数据来训练深度神经网络，这催生了通过量产车“影子模式”等创新采集手段，旨在自动化捕获人类驾驶员与算法判断不一致的长尾场景，相应地，数据融合的核心从“数据级对齐”跃升至“特征级融合”，以 BEV（鸟瞰图）感知范式为代表，通过神经网络将多摄像头、多模态信息在统一的俯视图空间下编码，合成出一张富含语义、几何与速度信息的动态特征地图。与此同时，合成数据技术开始服务于生成稀缺的关键场景特征，如极端天气下的物体形态或复杂的交通参与者交互，用以专项提升模型的鲁棒性。

2021 年特斯拉引入 BEV 后引发了业界的广泛关注，比如，理想汽车在 2022 年即做了 BEV 环视 ADAS 视觉算法与激光雷达感知数据进行前融合，并与与清华大学、MIT 合作完成了全球首个实时构建高精地图的公开工作。蔚来汽车正式推送 NOP+ 增强领航辅助功能，从统一框架进行功能设计，NOP+ 的综合通行成功率较上一代提升 40%。小鹏汽车 G9 车型获得广州智能网联汽车道路测试牌照，完全采用量产车为载体，大幅降低了 Robotaxi 的综合成本，促进了辅助驾驶与自动驾驶双向互哺，实现数据和技术能力的闭环迭代。

当前，面向端到端自动驾驶与大模型范式的探索，系统追求一个或少数几个能够直接映射传感器输入到控制指令，或进行复杂时空推理的通用模型。这就要求数据体系必须提供能支撑“感知 – 决策”联合优化与“世界模型”训练的连贯时空序列。因此，采集的重点从静态的标注帧转向大规模、时序化的真实驾驶视频流。这种合成数据需要具有空间一致性、时间连贯性及物理可交互性等特征，为训练具备预测、规划和因果推理能力的通用智能体提供了至关重要的数据基础。

在时序化视频流采集方面，蔚来汽车在蔚来世界模型 NWM（NIO WorldModel）中主张群体智能与生成式仿真，2026 年 1 月正式推送的蔚来世界模型 NWM 全新版本，首次在国内将完整的闭环强化学习技术深度融入智能驾驶研发体系。依托长时序环境推理能力与高频次闭环训练机制，模型可自主理解道路动态、交通常识及空间关系，显著降低对人工标注数据的依赖。理想汽车在 2026 年初明确提出了从“2D ViT”向“3D ViT”的架构跃迁，理想 VLA 司机大模型的训练数据已超过 1000 万 Clips。小鹏汽车的第二代 VLA 智能驾驶系统同样依赖覆盖 1 亿段驾驶视频的训练数据（等效人类驾驶 6.5 万年），实现视觉信号到驾驶指令的直接生成。

在空间一致性、时间连贯性及物理可交互性的数据生成方面，文远知行的 GENESIS 世界模型平台实现了突破。该系统不仅能高保真复刻真实路况，还能自动生成极端场景、自动诊断算法弱点，其生成数据与真实世界分布偏差小于 5%，实现“训练即实战”。地平线与 iCAR 联合发布的 HSD（Horizon SuperDrive）一段式端到端方案，通过结合强化学习算

法与 VLM 大模型“通识外挂”，在不依赖高精地图的情况下，实现对复杂道路要素的理解与迁移，能够在“未见过的场景”中快速理解环境并生成应对策略。

人形机器人与环境的交互方式更为复杂，涉及非结构化环境、全身动作控制与精细操作，这些包含物理常识的高维数据，获取和使用的成本高昂，导致其研发难度可能高于自动驾驶。自动驾驶工程化中不同功能层需求不同，具身智能的高层规划、技能模仿、底层控制也具有类似的差异化的数据供给。例如，高层任务规划需要包含任务逻辑与对象关系的视觉数据，操作任务的模仿学习需要高保真的动作轨迹数据等。自动驾驶中 BEV 特征地图的成功表明，将多模态原始数据融合并提炼成一种紧凑、结构化、适合决策的中间表征，远比直接处理原始信号更高效。然而，具身智能使用更多的专用传感器，意味着更复杂的标注工作，这些都为数据处理带来艰巨的挑战。

3.3 数据驱动的闭环：仿真优先，真机验证

自动驾驶系统的工程化落地中，数据通过采集、标注、训练、测试、回传等环节，持续驱动算法和系统优化，构建起了一个高效、可靠且可迭代的数据驱动闭环。这一闭环并非直接依赖于实车海量路测，而是遵循“仿真优先，真机验证”的核心原则。

“仿真优先”源于对研发确定性与经济性的极致追求。其核心价值首先体现在风险前置与成本控制上，在软硬件集成初期，将算法置于高保真仿真环境中验证，可以隔离真实车辆动力学、传感器噪声等复杂变量的干扰，在虚拟空间快速暴露算法逻辑与接口问题，避免将不稳定代码直接部署于昂贵且危险的实车平台。

仿真提供了确定性的测试与无限场景覆盖能力。任何交通场景，尤其是极端、危险的长尾场景，都可在仿真中被确定性地创造、精确复现和反复测试，实现了“过程可重复、结果可预期、问题可追溯”的工程理想。此外，这带来了闭环迭代的速度与规模革命。云端并行仿真可在数小时内完成相当于数百万公里路测的场景覆盖，实现算法版本的快速迭代与验证，形成了研发效率的数量级优势。纯粹依赖真机路测来积累里程、发现和解决长尾问题，在时间、成本和安全上均是不可承受之重。

在自动驾驶数据闭环中，仿真贯穿于数据采集、处理、应用的全流程，是驱动闭环运转的关键技术要素，扮演着三重核心作用。

第一，它是带精确真值的“数据工厂”，能自动生成无限量的、带有像素级或物体级精确标注（真值）的多模态数据，为感知等模型的监督学习提供了至关重要的初始燃料和扩充能力，并能按需提供从原始信号到高级语义的任一层次数据。

第二，它能串联多层次研发闭环。通过与不同实体组件结合，仿真支撑了从模型在环（MiL）、软件在环（SiL）到硬件在环（HiL）的递进式验证体系，确保从算法到产品的集成平滑可控。当在实车测试中发现故障时，仿真还能用于精准回溯，在虚拟环境中复现问题以定位根因。

第三，它是应对长尾场景的核心武器。通过将基准场景参数化，仿真能生成海量 corner case 进行穷举或抽样测试。更重要的是，它能与量产车的“影子模式”连接，构成完整迭代环，“影子模式”在真实世界中发现未知场景，仿真平台则负责将其重建、参数化泛化，用于算法修复与回归测试，最终通过 OTA 完成闭环。

自动驾驶探索出的“仿真优先，真机验证”范式，正深刻地指引着具身智能数据体系的走向。越来越多研究机构和企业开始采用“大规模仿真数据预训练 + 少量高质量真实数据微调”的混合训练模式，试图复现自动驾驶领域的成功路径。然而，一个根本性的差异在于：自动驾驶的“冷启动”相对容易，而具身智能的数据飞轮却面临着从零起步的“先有鸡还是先有蛋”困局。

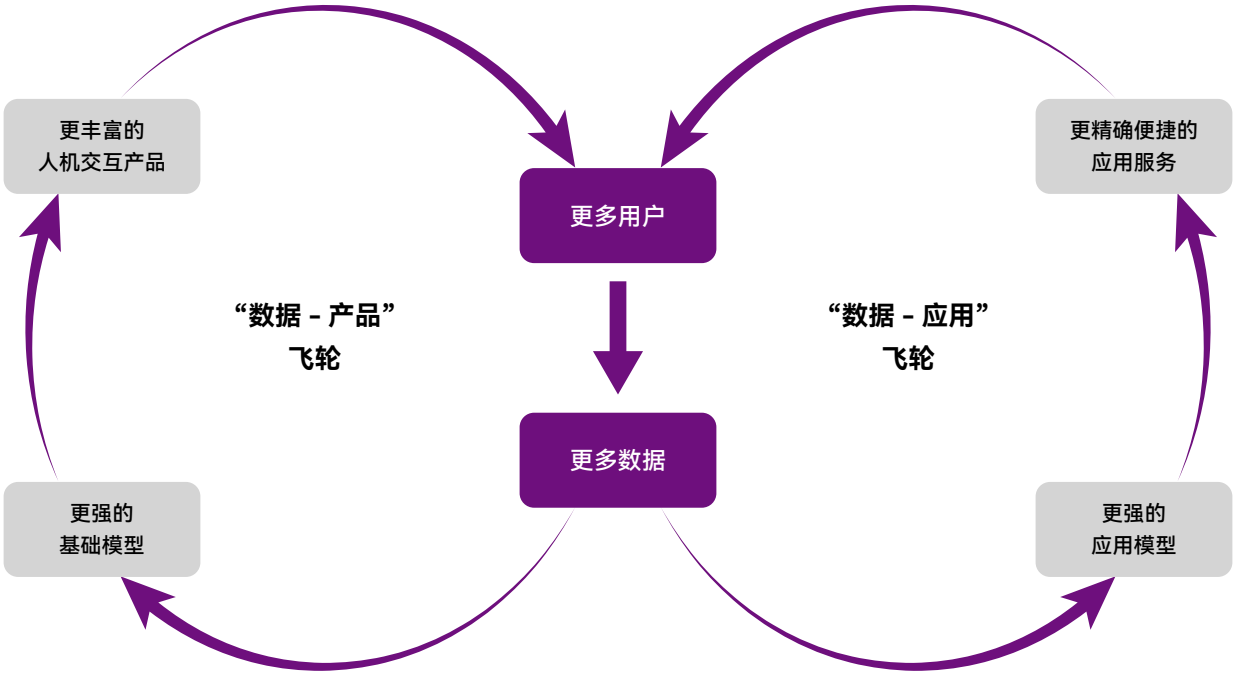


图 12　数据飞轮结构与迭代示意

车辆拥有明确的产品定义和既有的行驶场景，这使得自动驾驶的数据采集可以与产品销售同步启动——每卖出一台车，就新增一个数据采集节点，数据飞轮与产品规模自然共生。而具身智能截然不同，机器人还没有大规模进入真实场景，就没有足够的数据来训练智能；没有足够智能的机器人，就无法进入真实场景创造价值。数据飞轮在启动前是断裂的，需要企业主动、额外、持续地投入资源进行数据采集，无法像自动驾驶起步时“边卖车边采数”。

此外，具身智能必须搭建覆盖数据全生命周期的专业管理平台。因为行业从硬件体系到数据体系均缺乏统一的标准，没有类似自动驾驶的 3D 融合标注体系，直接导致无法评判仿真环境生成的数据价值。

本章从自动驾驶产业的发展历史出发，分析总结自动驾驶数据体系演变过程中的经验与教训，论证了单一数据采集路径的不足，以及智能的泛化难以通过静态真实数据的堆砌来实现。自动驾驶从实验室到商业落地的转变，源于构建了一个动态的、算法与数据共进的混合生态系统，以真实数据为锚点和校准基准，以合成与仿真数据为规模化扩展和加速迭代的核心引擎，这种范式为具身智能数据路线的选择提供了重要参考。

04 具身智能数据发展评估

自动驾驶对规模化与商业化的核心诉求，驱动其数据体系从纯真机采集的静态数据转向仿真与真机数据混合的动态生成。

这一过程揭示了数据与算法相互依存的基本规律：在任何算法应用的前期，数据的有效性都非常显著；而当算法效果提升到一定程度，需要转向寻找对算法改进有效的“高价值、特异性”数据。

因此，具身智能的数据采集不应是盲目的，而应针对模型中亟待加强的特定能力层进行定向补充。要么增加数据规模，通过样本量级获取特殊数据，要么选择数据挖掘，提升获取特殊数据的概率。

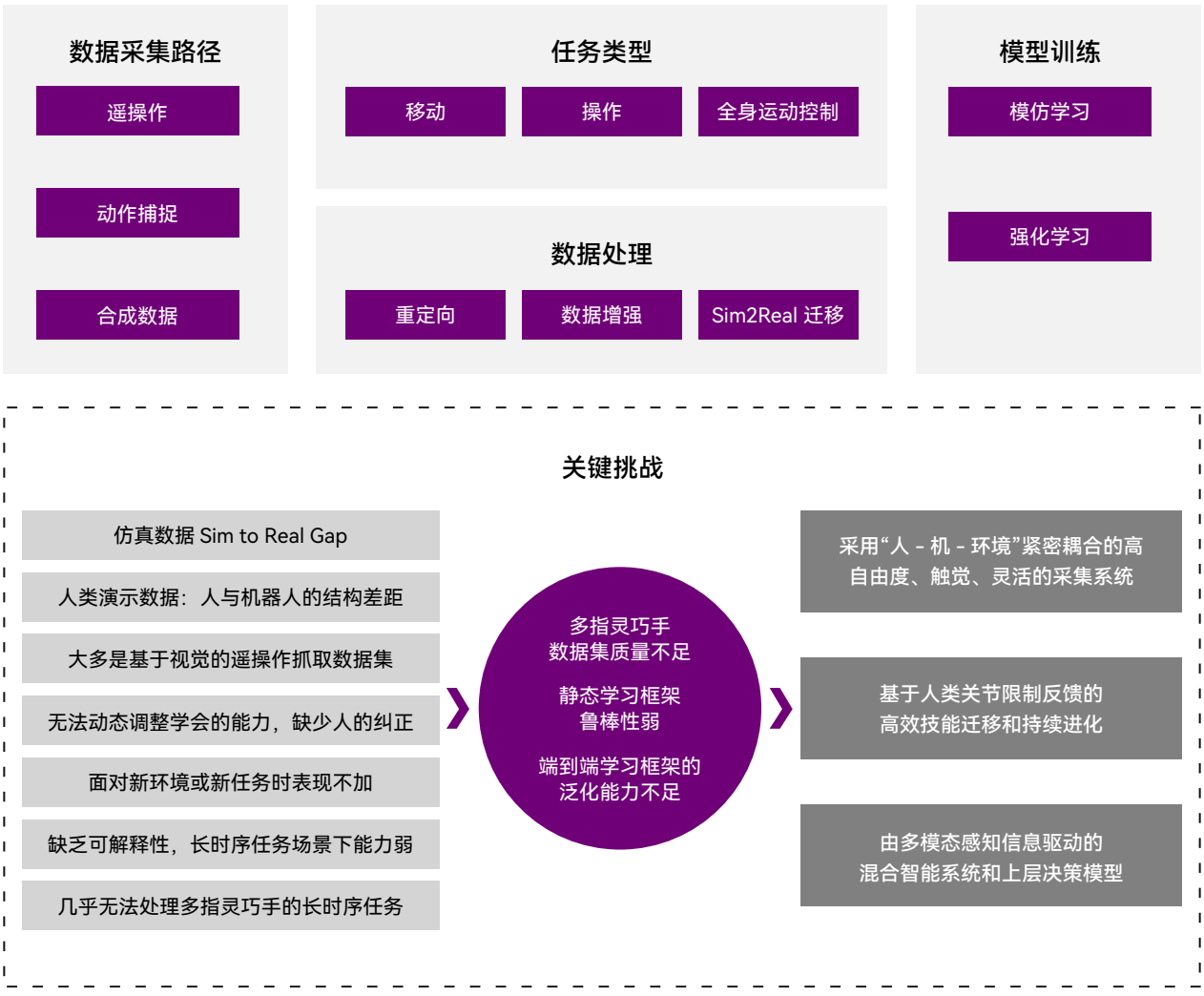


图 13 具身智能数据框架及挑战

4.1 真机遥操作数据在不同发展阶段提供不同价值

真机遥操作数据能同步记录下多维度的物理状态，提供了高保真、因果明确的物理交互轨迹，这种蕴含真实物理属性的载体，在具身智能发展过程中具有不可跨越的地位，是当前具身智能数据的黄金标准。

具身智能当前发展类似于自动驾驶的模块化探索阶段，研究团队分别致力于步态控制、抓取操作、视觉感知等独立模块的突破，这与自动驾驶早期分别改进传感器、算法与底盘的发展阶段存在一定相似，是必要的基础积累。真机遥操作的数据价值在于构建了高质量的动作 – 状态配对数据集，直接驱动行为克隆等模仿学习算法，使机器人快速获得可用的基础技能。此时，数据的有效性极高，每一条演示都能显著提升策略在特定任务上的成功率。

随着 VLA 模型的应用与效能提升，真机遥操作数据为代表的真实数据采集价值无法恒定，数据价值和使用定位因此会发生明显变化。作为关键的真值来源和性能上限，其固有的局限也开始成为模型性能持续提升的瓶颈。

首先是泛化能力局限，通过昂贵人力采集的数据，天然集中于有限场景和物体。由此训练出的策略，在面对新物体、新环境布局或需要多步骤推理的长时序任务时，泛化能力迅速衰减。模型学到的往往是特定场景下的动作记忆，而非可迁移的物理常识与推理能力。

然后，必须面对规模与成本的硬约束。与自动驾驶通过“影子模式”实现数据规模化的路径不同，机器人遥操作高度依赖专家，成本高昂、效率低下，无法实现数据量的指数级增长，纯遥操作路径在经济和工程上均不可持续。

最后，是数据噪声与质量瓶颈，操作员的手动控制不可避免地引入抖动、延迟和主观偏差。这些因素在训练高水平策略时可能成为干扰，而筛选和清洗此类数据本身又构成了新的成本。同时，操作员的演示难以提供超越人类水平的、更优或更具创新性的解决方案。

真机遥操作数据能够作为启动和校准智能体学习的黄金标准，但其自身难以作为单一数据源驱动模型实现通用泛化。当前多条数据采集的路线也展现出，大规模预训练的主体需要更多元、更经济的数据，而在最终校准、细化和验证等环节，遥操作数据仍有不可替代的高价值。

4.2 无本体数据采集有望推动模型性能

无本体数据采集的兴起，源于破解数据“成本 – 规模 – 多样性”不可能三角的迫切需求。2024 年斯坦福大学提出 UMI 框架，开启了无本体数据采集的新思路，希望通过技术手段，将人类的操作智能与物理环境进行解耦和抽象，形成一种可被规模化采集、且能跨机器人本体迁移的中间态数据，从而为模型提供近乎无限的预训练燃料和物理先验。

当前，无本体数据采集主要沿两条子路径协同推进，它们在模型训练中扮演着分工互补的角色。一条路径是带有轻量化传感设备的中低精度人类动作捕捉，可以看作是 UMI 路线的直接延续和工程化深化，UMI 系列设备以夹爪为主要形态，侧重于末端位姿采集，而动作捕捉技术和设备的应用能够很好弥补硬件形态带来的数据维度和自由度问题，区别于传统光惯类动作捕捉的专业设施和高昂价格，轻量化采集设备的穿戴体验更好，对人体正常运动操作干扰较小，同时还能增加触觉等重要感知维度。

当前，Generalist AI 不再围绕特定机器人本体建立昂贵的线下采集工场，而是让数据采集设备与机器人本体彻底解耦，其发布的 GEN–0 具身基础模型在 27 万小时人类操作视频数据上完成训练，首次在机器人领域验证了 Scaling Law 的存在。Sunday Robotics 团队花了数年时间迭代 100 次技能捕捉手套，在积累了百万级人类操作数据后，才正式启动机器人本体 Memo 的设计，在 2026 年初完成了一项被业界称为“里程碑式”的任务：从餐桌到洗碗机的完整家务链。在单次自主运行中，Memo 完成了 33 次独特灵巧互动、68 次总互动事件、跨越超过 130 英尺的自主导航，整个过程无需重置、无需遥操作介入。国内初创公司无界智航已开始推出相关产品原型，以低成本的触觉传感设备为支撑点，跟踪顶级模型算法的前沿动态，结合仿真数据的能力，有望提供大量物理可行、成本可控的操作轨迹和技能原型。

另一条路径更加极致，从人体运动视频中提取数据，但数据精度更低，处理难度更大。海量的视频数据对具身智能的潜在意义，类似于互联网文本数据对大语言模型预训练的作用，从高层的任务理解逐渐向中层的动作规划渗透，更多为模型提供物体运动、空间关系变化等深度推理能力，如果进一步实现执行控制，仍需要前一路径中采集所得的操作轨迹进行优化。

相比于高保真的遥操作数据，无本体数据在动作完整性（如缺乏精确的力触觉信号）和运动精度（存在视觉估计误差、本体映射偏差） 方面存在天然不足，其带来的工程难度不容忽视。

第一，数据治理层面，需要构建高质量数据管线，将硬件标准化、采集 SOP、实时质检、数据清洗标注无缝集成，技术和管理复杂度高。

第二，算法设计层面，核心是利用算法先验来弥补数据信号的缺失。源于人体动作的数据几乎都需要使用运动重定向（Retargeting）方法，需要针对数据特性（噪声模式、缺失模态）定制模型架构和训练策略，如设计专门的适应器（Adapter）或修补（Imputation）模块，无法直接套用现成方案。算法与数据的紧密耦合考验团队对算法的理解程度。

第三，建立工业级的数据质量评估体系，由于缺乏绝对真值，需要依赖分布一致性等间接指标，并结合最终的下游任务性能来综合评估数据有效性。

无本体数据采集开辟了一条在规模和多样性上拥有明显优势的新路径，其与高精度真机数据的关系并非对立，相互协同有望共同构成通用智能训练的混合数据生态。

4.3 仿真系统是一套必要强大的非完美工具

在自动驾驶领域，仿真系统是一个由多层次、模块化工具链耦合而成的工程体系，贯穿整个产品生命周期，以服务于特定任务的辅助计算。

以 NVIDIA 的 DRIVE Sim 平台为例，其底层是 Omniverse 提供的物理准确渲染与动力学基础；其上则构建了从传感器仿真、场景生成、到车辆高保真动力学模型、再到测试管理与结果分析的全套工具。通过设定虚拟环境，将“试错”这一智能诞生过程中的必要但高成本环节，变得规模化、安全化和经济化。

当下，从结构化的道路转向开放、充满非结构化交互的现实世界时，具身智能对仿真技术提出了远高于自动驾驶的要求，传统物理仿真模型的局限性被急剧放大，模型保真度与交互复杂性的矛盾成为核心焦点。一方面，具身智能的操

作涉及软体形变、复杂摩擦、细颗粒物学等，传统刚体物理引擎难以精确模拟，传统物理模型存在固有简化，导致仿真策略在真实世界执行时，因微妙的物理差异而失败。另一方面，相对道路这个有限语义的场景空间，具身智能面对的场景和任务可以看作是无限的，依赖人工进行 3D 建模和场景编排，无法满足对海量、多样化训练数据的需求。在这种限制之下，即使使用强化学习方法，智能体习得也只是在特定虚拟物理参数下具体策略，难以实现能力的泛化。

为满足具身智能的数据和训练需求，模拟器在科研领域也开始出现新的技术创新。比如 Genesis、DiffSim、PlasticineLab、FluidLab 等可微分物理引擎，能计算物理过程的梯度，而不再是传统物理引擎中的搜索，从正面解决传统物理模型的精度问题。从另一角度，基于学习的动力学模型（Learning-Based Dynamics Models）可以模仿人类的"直觉物理学"，通过从环境中学习，构建一个能够预测动作对未来状态影响的模型，通过神经网络直接从数据中学习动态规律，无需显式物理公式，更适合复杂、信息不全的真实场景。但这两种路径都处于前沿探索中，技术门槛高，相应的开发人员少，产业落地的周期不明确。

具身智能的仿真环境搭建是一条极具战略潜力但当前仍充满工程坎坷的必经之路。当前，科研领域开始探索生成式仿真，希望构建一个从任务提出、场景生成、到训练监督全自动化的数据闭环工具链，企图在数字世界中为具身智能复刻一个完美世界。仅仅是将多个功能模块无缝衔接集成，其系统工程的复杂度已经远超过单一算法的创新。

同时，完美仿真工具存在着隐形成本高昂和商业化能力弱的风险。虽然仿真环境可以批量生成数据，在单条数据的成本上明显优于真实采集，但是这并未计算前期研发过程中的顶尖人才投入、研发及生产过程中的算力配套等成本。自动驾驶工程化实践表明，即使出现了这样的完美系统，对于大多数企业和开发者来说，经济性至关重要，相比于一个昂贵的完美系统授权，不如为研发团队配备多种工具的多个席位效率高。

本章节从自动驾驶的发展经验出发，跟据当前具身智能的发展阶段，重新评估了目前三种主要的数据采集方式，本文认为三种数据路径并非替代关系，因其各有优势，在评估数据价值时，应在系统工程的角度上，考量模型能力瓶颈的突破效率，并因此整合调度数据资源投入，建立起容纳多种数据采集路径的基础设施和闭环体系。

05 数据视角下的渐进式商业化道路

具身智能还未迎来“GPT3.5”时刻。与自然语言处理领域“模型即产品”的线性演进不同，具身智能的链条更长、约束更多、变量更复杂。当 VLA 模型的参数规模从 7B 扩张到更大的量级后，能力增长并未如预期般持续涌现，反而呈现出了“边际递减”的上限。然而，这种能力上限并非单纯由模型架构决定。a16z 的深度洞察指出，实验室里 95% 成功率的策略，一旦进入真实仓库，光照、背景、视角、物体材质发生变化，成功率可能迅速跌到 60%。

从数据视角出发，数据瓶颈提供了一个确定且渐进演化的硬约束。自然语言处理之所以能够实现快速跃迁，关键在于它解决了自监督预训练的问题，能够压缩海量互联网知识。而在具身智能中，行业对视觉的编码方式和 3D 空间的推理机制等问题仍未形成统一认知。高质量真实数据的稀缺、多模态融合的工程难度、仿真到现实的迁移鸿沟，共同影响了智能体的能力边界与商业化场景的广度和深度。

因此，本章将从数据规模与质量的视角出发，尝试推演具身智能商业化演进的关键阶段，并勾勒各阶段的标志性特征。

5.1 少量数据构建原型和工程环境的执行能力

在商业化起步阶段，侧重快速打造可展示、可部署的最小可行产品，以最低成本和最快速度，证明特定技术在受限环境中的工程化可行性。此阶段的数据积累少，目标是获取易得、专精的数据，目前业内通常采用专家遥操作或提前预设的程序，高效利用数十至数百条高质量演示数据，训练机器人掌握一个或一组高度结构化的确定性子任务。

当原型验证了单点技术的可行性后，商业化需要考虑工程环境的执行能力。正如当前行业内大量公司，在完成原型产品后，目标进入工业制造领域的某个环节，虽然任务范围相对固定、执行路径清晰，但是对精度、可靠性和节拍要求极高。这些要求导致具身智能只能面向某些对节拍要求不高的离散产线场景，例如，在 3C 电子装配线上完成芯片贴装、螺丝锁付等，这类场景环境可控、动作规范，追求的是专业设备满足精准稳定、耐疲劳的要求。

当前发展阶段内，商业化价值的矛盾在于具身智能产品功能定义不清晰，导致成本高昂，同时，相比传统的工业机器人，在场景中替代高重复性的人工环节稳定性不足，商业竞争力略显不足。具身智能产品的投入产出比（ROI）取决于任务价值、替换人力的成本及部署的稳定性，例如，典型的工业客户会进行精准的 ROI 计算，若替代一名年成本 10 万元的工人，则对应机器人的设备成本加上维护费用，通常需在 1.5 至 2 年内收回，这要求具身智能产品的售价需降至 10–15 万元级别，同时需要满足几乎类人的工作效率和 7x24 小时连续作业的无故障要求。具身智能系统的复杂性，尚无法与经过数十年优化的专用设备相比，导致客户买得起，但用不起。

少量数据的发展阶段，更多是具身智能企业内部能力的构建。不仅考验团队的算法在特定任务上的性能，更是整个团队定义问题、设计数据流水线、进行快速实验和工程集成的综合能力。如何避开已被传统机器人占用的领域，精准定位那些需要一定的感知灵活性、轻量级的移动能力以及简单非刚性操作的场景，才是进入商业化成本与稳定性攻坚的基础。

5.2 聚焦场景，大量数据驱动算法迭代与标准化

参考自动驾驶的发展，不同场景在基础技术层面几乎没有本质区别，差异主要在于数据与训练量，技术应用根据环境要素的复杂程度划分为不同细分场景，这无形中也对人工智能技术进行了切分。反映在具身智能领域，在有限资源下聚焦场景，实现数据规模与能力表现的最佳平衡是关键课题。

目前行业正在努力构建多种场景，并广泛采集数据。基于遥操作采集方式构建场景数据采集，投资巨大、回报周期长、且直接商业收益不明确，使得单纯依靠商业公司难以独立完成。在此背景下，国内各地政府积极支持和参与，比如北京石景山区的国内最大人形机器人训练场、上海国家地方共建人形机器人创新中心打造的“麒麟”具身智能训练场、天津的帕西尼具身智能超级数据工厂，以及杭州、成都、宁波等地均有相关人形机器人试验场落地，据不完全统计，国内已建成或计划在建的具身智能训练场达到 20 余家，其中公开披露的 10 家训练场总面积超过 4 万平方米。

然而，依赖遥操作采集的数据与特定本体结构强绑定，原本定位为公共产品和基础设施属性的数采训练场，在推动产业发展上遇到了数据孤岛问题。

更深层次的意义上，是行业标准化的缺失。正如 ImageNet 催生了计算机视觉的复兴，具身智能领域亟需权威的、公认的基准测试数据集与评测环境。这些数据集应包含一系列具有代表性的任务，并提供标准的真值数据。它们将成为衡量不同算法和机器人平台性能的标尺，让产业竞争建立在客观、可比较的性能指标之上。如果缺失基准数据，导致新模型的效果提升会是单方面的，显然会出现既是裁判又是球员的现象，一旦模型部署到产品端就会不断出现问题，无法实现真正的技术落地。

5.3 海量数据实现高阶功能的闭环拓展

当前两个阶段在多个垂直领域得到验证后，标准化的确立和技术路径的收敛将推动产业进入新的发展阶段。具身智能的终极愿景是迈向具备跨场景任务理解、自主复杂规划与终身学习能力的通用智能体，这需要突破特定场景的数据分布，让模型接触到物理世界和人类意图的长尾多样性，这需要构建起一套动态更新、实时反馈的多模态数据处理闭环。

未来，伴随“云 – 边 – 端”协同技术架构的成熟，将可能实现机器人能力的解耦与重组。参考语言大模型的服务模式，未来在云端将出现智能孵化与调度中心。按照当前算力的发展趋势，云端将利用大规模算力，基于来自全球机器人反馈的数据，在数字孪生环境中进行持续技能训练、测试与优化。同时，云端维护着一个庞大的、标准化的技能库，每个技能都是一个经过严格验证、可独立部署的模型包，如同一个庞大的应用商店提供调用和下载。云端根据用户的任务请求，动态调配和组合技能，形成复杂任务的解决方案，并选择最优的机器人或机器人集群去执行。

由于机器人工作的室内场景，工作环境更为隐私，需要边缘侧服务器承担本地机器人的实时协同、数据预处理和隐私敏感计算，确保在断网或高延迟情况下的基础功能运行。

机器人本体进化成为一个相对标准化的通用移动计算平台，它承载着基础的感知、定位、导航和运动控制能力，并通过标准接口从云端或边缘端按需加载和执行技能包、环境信息等。

因为高阶功能的拓展，可能衍生出“智能即服务”的商业模式。当前，具身智能行业仍采用传统的机器人行业一次性售卖的商业模式，需要客户支付高昂的一次性费用购买机器人本体，供应商的利润主要来自硬件差价和有限的后期维护。这种模式存在两方面问题，一是高昂的资本支出阻碍了中小企业和个人用户的普及，二是面临价值固化问题，机器人出厂时的能力即为其能力上限。

从电脑到手机，行业竞争焦点从“功能的堆砌”转向“平台的可靠性、成本与通用性”。具身智能也可能走向类似路径，未来硬件本体标准化，利润变薄但市场总量急剧扩大，一个全新的、庞大的开发者群体将应运而生。用户可以根据自身需求，像在应用商店购买软件一样，订阅单个或多个技能。这一商业范式的普及，将催生一个全新的、更加精细分工的产业生态。同时，商业模式的转变不仅将扩大市场空间，更将重新定义人、机器人与智能之间的关系，推动社会生产生活方式的深刻变革。

综上，具身智能的商业化是一场与数据深度绑定的马拉松。通向高阶智能的发展路径仍是渐进的，先通过小数据在确定性问题中证明工程价值、获取初始现金流；再通过深耕场景，用大量数据驱动迭代，建立垂直行业壁垒；最终，海量数据与前沿算法相结合，不断提升智能表现水平。对于创业公司而言，生存与壮大的关键在于精准识别自身所处阶段，并构建与之匹配的数据获取与转化能力，选择出最现实的商业化图景。

06 机会与风险总结

当前，具身智能产业处于爆发前夜的关键阶段，高质量、多模态的交互数据，已成为驱动具身智能产业发展的核心要素。一方面，大模型的进展为机器人赋予了更强的认知与推理能力，使其处理开放世界任务的潜力增加；另一方面，高昂的数据采集成本与规模化数据供给不足，制约着模型能力的充分释放与快速迭代。这一矛盾正在创造巨大市场机会。本章将基于前文对技术趋势与商业路径的分析，总结具身智能数据领域的投资机会与潜在风险，旨在为行业发展提供前瞻性与实操性兼顾的参考。

6.1 发展机会分析

从数据视角看，具身智能的数据使用仍停留在早期阶段，大量的感知数据仍未被开发利用，因此也产生新的市场机会。本章节围绕“如何高效地获取、处理和利用数据”的逻辑展开，希望对数据这一支撑产业迭代的关键基础设施发展提供助力。

6.1.1 感知技术创新，为多模态数据提供入口

传感器不仅是数据的源头，其内置的预处理能力正决定后续模型能获取怎样的物理世界信息。从“被动采集”到“感算一体”，感知层正在直接影响智能的上限，此时不仅应该关注硬件层面的创新，更应聚焦于集成了算法和初级处理能力的智能感知模块。

例如，AI 视觉应用过程中，许多芯片集成了人脸识别算法，能直接在芯片上完成特征提取。这类智能硬件能极大减轻中央处理器的负担，降低系统延迟和功耗，类似的能实时输出接触力矢量和材质估计的触觉传感器，也存在巨大的应用潜力。

专注柔性技术的企业正努力在性能、成本和规模化之间寻找平衡点，同时尝试提供模型能力，比如他山科技不仅提供自研传感器，也提供基于多传感器的分布式协同感知与控制模型，使机器人能够整合多个触觉单元的信息，实现更智能、更灵巧的操作。另有如模量科技、赛感科技、悟通感控等企业从不同技术路线提供智能传感设备。

这对其下游的末端执行器及数据采集设备也具有重要价值。单独的硬件产品仍无法解决具身智能的发展困境，需要能够将视觉、触觉、力觉等数据进行硬件级同步采集和融合的平台，才能为训练更全面的具身模型提供高质量的素材。部分具身智能企业正致力于打造覆盖“指尖 – 关节 – 全身”的立体化力触觉感知网络，将灵巧手的控制形成一个独立模型，比如帕西尼感知科技专注于多维触觉传感技术，其发布的多指触觉灵巧手 DexH13 集成了近 2000 颗自研高精度触觉传感器，能实现压感、摩擦、软硬质地等 15 种多维触觉感知，公司已构建从核心传感器、灵巧手到人形机器人整机的全链条产品能力；傅利叶智能 GR–3 机器人通过在头部和躯干集成 31 个触觉传感器，使机器人能够感知和响应人类的触摸，极大地增强了人机交互的自然度和亲和力等。

6.1.2 数据采集与治理是推动具身智能走向标准化的底层基建

建立覆盖采集、清洗、标注、存储的全生命周期管理体系直接决定研发效率。具身智能的多模态数据采集、标注极其复杂耗时，传统人工标注方式成本高昂且效率低下，严重拖慢模型迭代速度。简智机器人的实践显示，实现“采集完成后 2 小时内新鲜数据送达模型”的目标，需要从硬件压缩、传输优化到云端处理的系统性能力。这种“全生命周期管理”能力的建设，是一项需要长期投入的工程基础设施。

当下发展过程中，已有大量企业提供与硬件相关的数据采集方案，但是数据标准的缺失，需要统一的数据采集平台，实现超越商业利益的意义和责任。围绕行业发展的数据难题，数据采集与治理服务的定位越发清晰。

第一，面向产业共性的基础设施，大幅降低创新成本投入；

第二，助力突破灵巧操作数据采集与本体解耦关键技术，能够打造高价值开源数据集；

第三，构建高保真、柔性设计的垂直场景集群，极大降低企业产品落地的迁移成本；

第四，建立独立权威的第三方评测体系，为重大决策提供科学依据，推动建立行业标准。

投资建设的数据采集与治理服务平台，为行业提供低成本、高保真的测试环境，能够主导基准数据集的采集与发布，利用其公信力和中立性，组织采集覆盖广泛场景、经过严格标注的基准数据集，并向学术界和产业界开放。更重要的是，政府可以通过平台，引导和鼓励企业、科研院所就关键标准达成共识。例如，组织行业标准制定工作组，推动硬件接口、数据格式等标准的建立，避免市场陷入低水平的重复竞争和生态锁定的内耗。

6.1.3 关注垂直场景解决方案，加速模型训练与部署

垂直场景是起点。聚焦具体行业、解决确定性问题、创造可计算的 ROI，是目前最具商业明确性的路径。

垂直场景为数据标注、模型架构和评估标准提供了天然的收敛框架。一个具体场景中，任务成功的标准变得明确、可量化，围绕场景企业能够建立专属的数据标注规范、场景特征库和性能基准。其次，为加速迭代，垂直场景解决方案需要构建与该场景深度绑定的模型开发与部署工具链，这并非通用平台，而是细分环节、高度定制化的生产力工具，类似于早期自动驾驶中的低阶功能模块开发。与自动驾驶不同的是，具身智能生态处于早期，合作更加开放，数据、算法、硬件方案的话语权相对均衡。

因此，投资垂直场景解决方案，实质上是投资其将具体行业知识转化为数据标准，再通过工程化工具实现规模化复制的能力。目前，在工业精密装配、仓储柔性物流、商业清洁等领域，这一路径已展现出清晰的商业化前景，但竞争较为激烈，而在具有更高商业价值的封闭、高危、长期有害等封闭场景上，还有待智能产品的进一步探索和开发。

6.1.4 真机失败数据正加速具身智能的落地进程

被忽视的负面样本对模型能力提升也有重要作用。在 demo 演示阶段需要成功演示学会“如何做”，需要使用筛选后干净的专家数据，但实际应用充满噪声，模型需要从大量失败案例中学会判断好坏，从而更好地自我进化。

当前，具身智能产业内已经开始实践探索。智元机器人提出了 ADC（对抗数据采集）模式，通过增加数据的信息密

度和多样性，以 20% 的数据量达到传统方案 2.7 倍的效果。同时，为配合人工采集方式，智元首席科学家罗剑岚团队提出了 HIL–SERL 系统，通过“Human–in–the–Loop”的强化学习，在后训练阶段，针对特定任务解决模仿学习的短板，让机器人在真实世界中 1–2.5 小时内学会多种高精度、灵巧的操作任务，成功率接近 100%。

6.1.5 世界模型是通往具身“GPT-3.5 时刻”的潜在路径，但仍需耐心

世界模型被认为是补足机器人“物理直觉”的关键拼图。智源研究院在《2026 十大 AI 技术趋势》中将世界模型定义为 AGI 新范式，指出行业共识正从语言模型转向能理解物理规律的多模态世界模型，“预测下一个状态”成为核心方向。

2026 年初，这一方向迎来密集突破。蚂蚁灵波发布的自回归视频 – 动作世界模型 LingBot–VA，首创“边推演、边行动”框架，使机器人能够像人一样在执行动作的同时预判环境变化，在 LIBERO 基准测试中任务成功率高达 98.5%。生数科技联合清华大学开源的 Motus，首次将 VLA、世界模型、视频生成模型等五种主流具身基础模型范式统一到同一框架中，在 50 项通用任务测试中，绝对成功率较国际顶尖的 Pi0.5 提升 35%。英伟达与斯坦福联合发布的 Cosmos Policy 则展示了基于规划的推理路径：模型先提出 N 个可能的动作系列，利用世界模型想象执行后的未来画面，再通过价值函数择优执行，在极具挑战性的任务中成功率提升 12.5%。

然而，世界模型仍处于早期探索阶段。当前模型在长时序预测中的误差累积、物理一致性保持、实时推理效率等方面仍面临共性挑战；“互联网数据 + 真实数据”路线与 Sim2Real 仿真路线正在并行探索，尚未形成统一收敛。技术突破需要时间，仍需要大量资金和时间投入。

6.1.6 数据路线之争远未终结，能否“完全无本体”仍是开放命题

不同于车辆和自动驾驶算法，具身智能的软硬件几乎在同步诞生和演化。这直接造成了具身智能数据飞轮的割裂，在产品数据和应用数据的两端出现目标函数的不一致。正如当前，具身智能企业追求第一性原理，坚持端到端模型学习物理世界交互，利用机器本体高保真的数据，使具身智能具备跨任务、跨场景、跨形态的天然泛化潜力。同时解决数据需求也迫在眉睫，无本体数据采集具有一定的规模和成本优势，亦有创新企业采用这类数据展现出优秀的 demo 能力。

未来或许不存在唯一正确的答案，只有场景适配。在通往通用具身智能的路上，路线之争远未终结，而这场争论本身，正是技术走向成熟的必经阶段。

6.2 风险与挑战

然而，具身智能领域也面临不容忽视的风险。技术的快速迭代意味着当前解决方案可能被更优路径替代；数据安全与伦理监管正在构建新的市场准入门槛；漫长的商业化周期考验着创业企业与资本的耐心；行业标准的缺失可能使早期创新受限于“数据孤岛”。因此，对所有行业参与者而言，这既是捕捉技术机遇的挑战，也是穿越产业周期的考验。

6.2.1 技术架构快速迭代与路径收敛风险

具身智能的技术栈远未达到稳定状态，核心挑战在于技术路径的不确定性。当前，主流的技术路线主要分为两大阵营：一是以“感知 – 规划 – 行动”为代表的模块化架构，其优势在于系统透明、易于调试；二是受大模型驱动的端到端架构，它直接将传感器数据映射为控制指令，潜力在于更强的泛化能力。然而，这两种路线孰优孰劣，尚无定论。更值得关注的是，可能出现第三种技术路径，比如基于世界模型的全新范式，能够通过内部模拟来预测行动后果，从而大幅减少真实数据的依赖。关注具备技术适应性与前瞻视野的团队，重点考察公司技术栈的灵活性，是否能快速适配不同的训练范式，灵活性比暂时的性能优势更为重要。

6.2.2 数据可用性验证的投入风险

数据的“可用性”验证本身，正在成为一项投入巨大、周期漫长且结果不确定的隐性成本。行业普遍认识到数据的稀缺性，却往往低估了“让数据真正可用”所需的系统性工程投入。数据的异构性与时空对齐难题，使得“可用”的标准极难达成，叠加数据标准的缺失，导致“可用性”难以被验证和复用。

数据质量的验证需要贯穿采集到标注的全链路投入。数据堂的实践表明，真正可用的数据需要经过多流程质检和专家团队的现场把控——在 4 万条有效操作记录的背后，是技术工程师对机械臂稳定性问题的反复克服、对任务多样性的持续调整，以及对轨迹复现的严格审核。这些投入并非一次性的，而是随着采集规模扩大而线性增长的持续性成本，可用性验证失败将对企业带来巨大沉没成本。

6.2.3 数据安全、隐私与伦理监管风险

数据安全与模型完整性面临全新挑战，隐私合规压力日益严峻。用于训练机器人的多模态数据集可能成为攻击目标，遭遇“数据投毒”，对于在物理世界中运行的机器人而言，此类安全事件已不再是虚拟损失，而是可能转化为真实世界的财产损害甚至人身伤害。同时，通过遥操作、仿生演示或环境扫描采集的数据，极易包含个人信息、商业秘密或受版权保护的内容。伦理与可控性成为监管焦点，最新的《人工智能安全治理框架》等文件明确要求公司的产品具备足够的透明性、可解释性和可控性。

6.2.4 产品功能安全保障缺失的人机交互风险

在评估具身智能企业的产品成熟度时，不能仅关注其功能演示的惊艳程度，更要审视其在安全保障上的系统性投入。具身智能在感知、决策、执行、交互各环节的失误，已不再只是数据层面的偏差，而会直接引发现实世界的物理伤害。

目前，物理层面的安全风险尚未形成有效的测评与防护体系。交互安全影响机理不清、测试方法及标准体系缺失、测试仪器与标定方法缺失，这意味着，当前行业尚缺乏一套公认的方法论来系统评估机器人在真实交互场景下的安全性。除此以外，在网络安全漏洞、认知诱导、环境中的视觉误导，都可能引发具身智能体决策问题，对交互对象造成直接伤害。

面对这些系统性风险，产业界未构建完善的安全防护体系前，产品的规模化都将存在巨大人机交互风险。

6.2.5 行业生态与标准缺失的风险

标准化进程的滞后将导致市场碎片化。如果机器人硬件、数据接口、技能表征长期处于不统一状态，那么数据工具公司就不得不为每个客户进行大量定制化开发，难以形成规模效应。这会导致研发资源分散，产品化程度低，企业长期陷在项目制的泥潭中，无法享受软件行业典型的高毛利和可扩展性。

6.2.6 商业化进程不及预期的风险

市场对具身智能的期望很高，但产业落地的步伐可能比乐观预测更为缓慢，创业者与投资者都需要面对可能较长的商业化周期。

当前，国内具身智能主流公司聚焦于三大市场：政府数据采集中心、高校科研设备采购、娱乐展演活动，当前行业内的多数订单仍属“试点验证”性质，仍然缺乏生产和服务性的商业化落地。行业落地的阶段性特征明显，真正实现跨场景的通用智能，可能需要十年以上的技术积累与生态培育。

市场需求的真实性与支付意愿需要冷静评估。当前，许多机器人应用场景的经济性尚未完全验证。衡量机器人经济性的核心标尺是投入产出比（ROI），而决定 ROI 的关键变量是机器人的投资回收周期与当地人力成本的比值关系。基于这一逻辑，海外市场成为更具商业确定性的优先选择。发达国家普遍面临人力成本高企、用工难、老龄化加剧等结构性矛盾，对自动化替代的需求更为迫切，市场接受度更高，定价空间和利润水平也更优。

站在 2026 年的当下回望，我们正在见证一场比互联网革命更为深远的变革。具身智能从来不是单纯的机器人产业，是人工智能从数字世界的“认知智能”向物理世界的“行动智能”的跨越。人工智能逐渐进入物理世界的浪潮，将是一个比互联网、移动互联网更具想象力的时代。从数据视角审视，通往这一未来的道路注定漫长而艰辛，这不会是一次 ChatGPT 式的突变，而是一场以五年、十年、二十年为尺度的渐进式演进。具身智能的数据领域，是一场关于未来智能基石的布局，最终，最大的赢家很可能不仅是技术的领先者，更是那些能深刻理解产业节奏、精准定位自身生态位，并能在复杂风险中构建起持续迭代能力和强大商业护城河的企业。

附录 常见数据集整理

当前具身智能数据集主要以操作演示为主，包含定义明确的操作任务开始，在真实世界或仿真环境中涉及相应的实验场景，利用传感器和动作捕捉技术采集机器人与环境交互的数据，经过标注和分析后，最终形成一个包含环境状态、机器人动作、物体属性和任务结果等信息的综合数据集。

表 2 常见具身智能操作数据集

数据集	数据形式	核心特点与定位	数据规模	数据模态	年份
BridgeData	真实	强调泛化能力，涵盖 71 个不同的厨房主题任务，分布于 10 个独特环境中，由 WidowX250 机械臂和 Oculus Quest2 采集	7200 个演示	彩色 – 深度图像	2021
Ego4D–Robotics	真实	由 Meta FAIR 团队与全球 14 所大学合作构建的大规模第一人称视频数据集	3670 小时真实世界第一视角视频，385 万条带时间戳的动作描述，涵盖 1772 个动词和 4336 个名词	RGB、音频、3D 环境网格、眼动追踪、多摄像头同步视图	2021
RT–1	真实	数据由 13 台机器人在 17 个月内采集，包含超过 700 中不同的语言指令，提供丰富的任务和物体种类	约 13 万个机器人演示	彩色图像	2022
ManiSkill2	仿真	包括 20 个任务家族、2000 多个对象模型，能够支持广泛的算法、视觉观察和控制器	400 多万个演示帧	点云 /RGB–D	2023
ARNOLD	仿真	专注于语言理解和连续目标状态学习，包含 8 种语言条件下的任务，涵盖 40 种物体和 20 个场景，基于 NVIDIA Isaac Sim 构建	10080 次演示	文本 /RGB–D	2023
Bi–DexHands	仿真	具有两只灵巧收的模拟器，提供 20 个手动操作任务和数千个目标物体	1638400 个步骤演示	力传感信息 / 点云 /RGB–D	2023
ROBOTURK	仿真	众包方式采集，旨在提供大规模标注数据，通过云平台混合智能手机作为运动控制器，让用户远程实时遥操在模拟环境中完成任务	超 2200 次成功的任务演示，137.5 小时的轨迹	视频流、运动传感、触觉	2023

数据集	数据形式	核心特点与定位	数据规模	数据模态	年份
BridgeData V2	真实	覆盖 24 个环境，支持通过目标图像或自然语言指令的任务调节，由人工操作和自主采集相结合	60096 条轨迹，50365 次远程操作演示，9731 次部署	RGB–D、音频、文本和触觉	2023
Open X–Embodiment	真实	汇集了 22 种不同机器人类型的数据，由 60 个单独的数据集组成，采用统一的 RLDS 数据格式	22 种机器人，超 100 万条轨迹，527 项技能	力传感信息 / 点云 /RGB–D	2023
RH20T	真实	以多模态、大规模、高多样性为特点，覆盖从简单抓取到复杂组装的多种场景	147 个任务，42 种技能，110000 个机器人操作序列	RGB、深度、双目红外、触觉、音频	2024
DROID	真实	大规模真机操作数据集，涵盖了 86 种不同任务类别和 564 个真实场景，由全球 18 个实验室在 12 个月内采集	7.6 万条演示轨迹，350 小时的交互数据	RGB–D	2024
ARIO	真实 / 仿真	包含了 300 多万条记录，结合真实世界数据和模拟数据，主要用于提高具身智能体的鲁棒性和适应性	258 个系列和 321064 个任务	RGB–D、音频、文本和触觉	2024
RoboMIND	真实 / 仿真	覆盖了家庭、厨房、工厂、办公、零售等场景，以及单臂、双臂、人形等多个形态的机器人，任务从基础操作到复杂的长时序任务，轨迹长度为 200–500 个时间步，由人工远程操作采集	5.5 万条机器人轨迹，279 个任务，61 个物体类别	文本 /RGB–D	2024
AgiBot World	真实	覆盖五大核心领域（家居、餐饮、工业、商超和办公），从基础操作到复杂交互的 80 余种日常生活技能，采集自智元自建的 4000 平实验基地，采集平台配备了 8 个摄像头、6 自由度灵巧手和全身 32 自由度的机器人	100 多个机器人的 100 多万条轨迹，5 个领域的 100 多个场景	RGB–D、触觉	2024
RealOmni–Open DataSet	真实	由简智机器人开源的大规模无本体具身数据集，涵盖 10 大场景任务，30+ 项技能，来自 3000+ 个真实家庭场景的自然操作	超过10,000小时，百万条以上（1Mil+ Clips）真实操作记录	超大 FOV 原始图像、高精度轨迹、语义标注、关节动作	2026

相较于具身智能的操作任务，运动控制发展起步较早，运动演示数据主要用于人形机器人，数据来源主要是人体姿态的动作捕捉、基于视频的人体姿态估计和合成数据。

表 3 常见具身智能运动数据集

数据集	数据形式	核心特点与定位	数据规模	数据模态	年份
Human3.6M	动捕	采用高精度动捕技术，由 11 名专业演员在受控环境中表演 15 种日常活动	360 万帧 3D 人体姿态数据	2D 和 3D 的骨骼关节位置、深度图像和视频序列	2014
KIT Motion–Language Dataset	动捕	整合多个运动捕捉数据库，并使统一的表示方法，便于数据处理	3911 个动作，6278 个自然语言注释	3D 的骨骼关节位置、文本	2016
AMASS	动捕	在共同的框架和参数化下统一了不同的光学标记，使用 SMPL 人体模型和 MoSh+ 方法，使人体模型和动作更逼真	300 多个主体和 11000 多个运动	3D 的骨骼关节位置	2019
HumanAct12	合成	源自极坐标图像和 3D 姿势数据集 PHSPD，通过时间裁剪和动作注释构建而成，包括 12 个主要动作类别和 34 个更细致的子类别	1191 个 3D 运动剪辑，共 90099 个姿态	3D 的骨骼关节位置	2020
HumanML3D	动捕 / 合成	结合了 HumanAct12 和 AMASS 数据集，包括日常活动和体育运动，由 5371 个不同的单词组成，动作总时长 28.59 小时	14616 个动作和 44970 个描述	3D 的骨骼关节位置、文本	2022
Humanoid–X	姿态估计	主要用于自然语言指令实现人形机器人的通用姿态控制，通过从互联网和学术数据集中挖掘视频，经过字幕生成、人体姿态估计、动作重定向等步骤创建	163800 个动作样本	视频、文本描述、3D人体姿态估计、人形机器人关键点和机器人动作序列	2024

在数据集整理过程中，本文发现目前研究工作尚少提供同一数据集下，测评不同机器人本体的性能及表现对比，推测可能是现阶段关注重心仍在于如何为特定本体有效利用数据，而非数据本身的通用性。然而，通过分析跨本体验证的案例和大规模数据集的组成分布，我们可以深刻理解不同形态的机器人与任务能力之间的映射关系。未来的数据采集及评测趋势，可能会随着像 Open X–Embodiment 这样的通用数据生态和标准化仿真基准（如 Isaac Gym）的成熟，而逐渐出现更系统的横向对比。

国际先进技术应用推进中心（深圳）

为技术对接场景，为企业对接技术

研究中心介绍

国际先进技术应用推进中心（深圳），简称国先中心，是在国家发展改革委、深圳市政府支持指导下设立，深圳市科技创新局作为业务主管部门，聚焦人工智能、具身智能、低空经济等重点领域，依托粤港澳大湾区数字经济研究院建设的具有技术先进性、平台开放性、资源国际化、运作市场化、服务专业化的世界级先进技术应用推广平台。

国际先进技术应用推进中心（深圳）

深圳福田区福保街道长富金茂大厦 1 号楼 57 层
电话：0755-83212983
邮箱：gxzx@idea.edu.cn
www.idea.edu.cn