

工业具身智能操作数据的采集、处理与可信共享

韦莎¹, 费思琪², 孙闯¹, 汤丰恺¹, 段泽明¹

(1. 中国信息通信研究院信息化与工业化融合研究所, 北京 100191; 2. 天津大学电气自动化与信息工程学院, 天津 300072)

摘要: 随着以智能机器人为代表的工业具身智能装备加速走向车间, 支撑策略学习、鲁棒规划与在线优化的高质量、多模态操作数据成为关键瓶颈。现有公开数据集多源于实验室或仿真环境, 在工艺约束、设备异构性、过程安全以及数据安全合规等方面与工业现场存在系统性差距。为此, 本文面向工业具身智能操作数据, 提出一套覆盖“数据类型—采集架构—处理流程—共享机制”的系统化建模框架, 梳理多源数据组织、关键处理环节及其工程化要点。为了进一步对齐《可信数据空间发展行动计划(2024—2028年)》中的数据可信管控、资源交互、价值共创核心能力, 本文基于可信数据空间系统构建了工业具身操作数据的参考框架: 以车间级具身智能单元作为数据节点, 通过边缘侧预处理、连接器、数据目录以及数字合约与使用控制等机制, 实现操作数据在跨主体流通利用中的可控授权、实时存证与结果追溯, 支撑可信共享与复用。最后, 本文结合工程案例给出该框架的落地路径, 为工业具身智能数据方案设计与可信数据空间标准化提供参考。

关键词: 工业具身智能, 操作数据, 数据采集与处理, 多模态数据集和可信数据空间

中图分类号: TP242.6

4—引言

4.1—工业具身智能与操作数据的重要性

“具身智能”强调智能体依托具体物理载体, 通过感知—决策—行动的闭环与环境持续交互, 从而涌现出更接近生物体的智能形式。这一思路可以追溯到 Pfeifer 等人在《Embodied Artificial Intelligence: Trends and Challenges》中的系统论述, 以及其著作《How the Body Shapes the Way We Think》中对“身体—环境—控制”耦合的强调^[1-4]。相比传统主要依赖互联网静态数据的“离身智能”, 具身智能将学习过程迁移到真实或高保真仿真环境, 通过连续的感知与动作序列来学习世界的动力学与任务结构, 被普遍视为迈向通用人工智能的重要路径之一^[5-6]。

近年来, 具身智能技术发展火热, 从具身机器人平台与仿真器, 到具身感知、具身交互、具身代理与仿真到真实迁移等核心任务均得到了较大的进展^[5-8]。尤其是在大模型驱动的具身智能研究中, 多模态感知与长时间尺度的交互轨迹已成为支撑具身智能体理解环境、规划任务和生成复杂动作策略的基础数据形态^[7-8]。在工业制造领域, 新一代智能机器人已经成为生产系统的关键执行单元。中国工业机器人市场规模持续增长, 机器人在汽车、电子、新能源等行业的装配、焊接、搬运和检测等环节广泛应用, 并被视为推动制造业转型升级和突破性创新的重要抓手^[9-11]。然而, 传统工业机器人多依赖人工示教、离线编程和局部优化控制策略, 难以在多品种、小批量、柔性制造等复杂环境下保持稳定的效率与质量。工业具身智能的引入, 使机器人需要从真实生产过程中的多源操作数据中持续学习和自适应, 以支撑

收稿日期: 2026-01-21; 修订日期: 2026-04-20

更高层次的自主决策与协同能力^[9,12]。

在算法和模型层面，多模态的基础模型正在快速渗透到机器人领域，大规模预训练模型通过在统一表征空间中融合视觉、语言、运动等多源信息，具备了跨任务迁移、语义理解和高层规划的潜力^[13-17]。以 Open X-Embodiment 数据集和 RT-X 模型为代表的工作，首次尝试在 20 余种机器人平台、数十万任务实例上进行统一训练，表明通过汇聚多机器人、多场景的操作数据，可以显著提升策略在新机器人和新任务上的迁移效果^[15]。目前，越来越多的专家学者认识到具身智能的能力上限越来越取决于可用操作数据的质量、规模及其覆盖的场景多样性，只有构建足够大规模、跨平台、跨场景的操作数据底座，才能真正释放工业具身智能与机器人基础模型的潜力。但是，与开放实验环境不同，工业场景中的操作数据往往同时承载工艺配方、设备能力、生产节拍、质量指标等关键信息，其采集、处理和使用必须兼顾安全性、合规性与可追溯性^[9-11]。因此，在工业具身智能背景下，操作数据不再只是“训练样本”，而是贯穿设计、仿真、上线、运维全生命周期的核心生产要素和战略资源。如何面向这些需求，系统性地设计工业具身智能操作数据的采集架构、预处理流程以及与可信数据空间相衔接的管理与共享机制，成为了具身智能在工业领域落地应用和大规模推广的关键。

1.2—现有机器人操作数据集及采集模式分析

围绕机器人操作任务，学术界近年构建了多种公开数据集与仿真基准。典型代表之一是 RL-Bench 基准^[18]，在统一的 Franka Panda 机械臂平台上设计了约 100 个视觉引导操作任务，提供多视角红绿蓝 - 深度图像 (Red Green Blue-Depth, RGB-D)、关节状态和动作轨迹等多模态观测，并通过运动规划自动生成大规模示教数据。与此同时，也有学者从任务维度系统梳理了机器人操作领域的数据集与基准，并首次从“数据采集—数据利用—泛化能力”视角提出统一分析框架，指出数据采集模式与数据集组成已经成为影响操作策略可迁移性和稳健性的关键瓶颈之一^[19]。

在真实世界多模态操作数据方面，DROID 与 RH20T 代表了当前“基础数据集”建设的前沿探索。DROID 数据集^[17]通过统一的开源硬件栈，在北美、欧洲和亚洲 18 家实验室分布式采集了约 7.6 万条操作示范、约 350 小时交互数据，覆盖 564 个场景和 86 个任务。RH20T 数据集^[20]则更侧重“接触丰富、传感多模态”的真实操作，包含 11 万余条操作序列和对应的人类演示视频，覆盖 140 余类任务与多款机器人平台，提供视觉、力/力矩、音频、关节力矩等多模态信息，旨在支撑基于示范的一次性泛化与复杂接触技能学习。

面向语言驱动长时序操作任务，CALVIN 基准^[21]在仿真环境中设计了长时间跨度的桌面操作任务序列，通过自然语言指令串联多个子任务，用于评估策略在零样本语言指令和新环境、新物体上的泛化能力。CALVIN、RL-Bench 等基准已经成为评估语言条件策略和视觉-语言-动作 (Vision-Language-Action, VLA) 模型的核心数据来源之一^[22-23]，推动机器人在“通过自然语言完成复杂操作任务”方向的快速发展。

然而，从“工业具身智能”的视角审视上述数据集与采集模式，仍可以看到很多局限：

一是场景与工艺鸿沟明显。现有数据集多聚焦桌面操作、日常物体或办公室/家庭等通用环境，其任务设计与真实生产线中的工艺流程、质量控制、生产节拍、安全互锁等工业约束之间仍存在显著差距^[17,19]。这使得直接基于此类数据训练的策略，在迁移到工业生产环境时往往需要大量再调试甚至重构。

二是缺乏“过程级”语义与数据治理信息。现有操作数据集在元数据设计上主要围绕机器人本体、任务标签和感知模态，对工业场景中关键的工艺参数、班组/班次信息、责任主体划分、数据权限与使用条款等描述相对不足^[19-20]。这不仅限制了数据在跨系统、跨组织间的可理解性和可追溯性，也难以满足工业场景对合规审计、责任界定和过程优化的需求。

三是尚未系统对接可信数据空间等数据基础设施。现有工作更多关注如何利用数据提升

策略性能，而较少在采集源头嵌入数据分类分级、脱敏与匿名化策略，也缺少在数据共享环节中对使用控制策略、审计日志和责任链路的统一建模^[19,21-23]。与此同时，国家数据局发布的《可信数据空间发展行动计划（2024—2028 年）》明确提出，要建设“可信可管、互联互通、价值共创”的数据空间体系，将可信数据空间作为联接多方主体、实现数据资源共享共用的新型数据基础设施^[12]。在这一政策目标下，现有机器人操作数据集在与行业数据空间、可信数据空间规则体系的衔接方面仍明显滞后。

综上所述，公开机器人操作数据集为具身智能研究提供了宝贵的“基础训练语料”，但从工业应用和国家数据空间建设的双重要求来看，其在场景贴合度、过程语义表达以及面向可信流通的数据治理能力上仍存在明显不足。

在工业物联网与智能制造快速发展的背景下，企业已经能够从可编程逻辑控制器（Programmable Logic Controller, PLC）、数控设备、机器人控制器、监控与数据采集系统（Supervisory Control and Data Acquisition, SCADA）、制造执行系统（Manufacturing Execution System, MES）、企业资源计划系统（Enterprise Resource Planning, ERP）等多层级系统持续采集大规模运行数据，但要把这些数据真正“用起来”，在工程落地层面仍然面临着多源异构与全链路集成难度高、数据质量不稳定、语义不一致、安全隐私与可靠性的系统性约束等一系列结构性挑战^[24-28]。

此外，要让工业具身智能操作数据真正“从车间走向数据空间”，还必须满足跨组织流通的合规与治理要求。一方面，需要与国内可信数据空间相关政策文件中提出的目标保持一致^[12]；另一方面，国际上以德国 Manufacturing-X 及其灯塔项目 Catena-X 为代表的制造业数据空间实践表明，要在跨企业、多层级供应链中实现数据主权、互操作和合规共享，必须在技术架构、治理模式和认证机制上形成系统设计，通过统一角色模型、核心服务、标准和认证流程来降低信任门槛和治理碎片化风险^[29-32]。

综合来看，现有具身智能与机器人基础模型研究已经充分揭示了多模态操作数据对于策略泛化和鲁棒性的核心作用，而代表性的开放数据集与基准虽然在任务数量、场景多样性和传感模态上取得了显著进展，但在与真实工业场景的工艺约束、数据治理要求和跨主体可信共享等方面，仍存在明显差距。因此，当前亟需在充分吸收现有数据集与基准经验的基础上，面向工业具身智能操作数据，构建既兼顾多模态学习需求、又满足可信数据空间治理要求的数据采集与抽象方法。

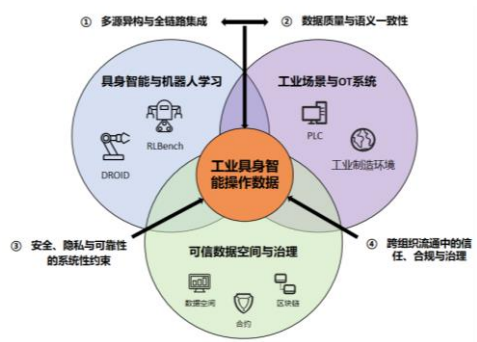


图 1 工业具身智能操作数据研究的核心挑战

Fig.1 Core challenges in industrial embodied intelligence operational data research

1.4 本文工作与贡献

本文从数据采集与处理的视角出发，尝试通过深度融合“具身智能—工业场景—可信数据空间”，搭建一条可实施复用的完整架构。本文的主要工作与贡献包括：

（1）提出面向工业具身智能操作数据的四维分类与建模框架。

结合具身智能和机器人操作领域的最新进展，本文从四个维度，对工业具身智能操作数据进行系统化抽象，该框架补足了现有具身智能综述和机器人基础模型综述中对“工业操作数据形态”的刻画不足。

（2）面向代表性操作数据集梳理并提炼可迁移到工业场景的采集与处理模式。

基于对 RL-Bench、DROID、CALVIN、RH20T 等代表性多模态操作数据集的系统文献调研与对比分析，本文从多方面进行横向比较，提炼出若干具有工程启发意义的模式，并分析其在迁移到工业生产线上面临的工艺节拍、安全约束、质量追溯等方面的约束。

（3）构建“工业具身智能操作数据→可信数据空间”的参考架构。

针对多源异构、数据质量、安全合规与跨组织治理等挑战，本文在不依赖特定厂商实现的前提下，提出一套从车间级具身智能单元到数据空间的参考架构，该架构一方面尊重工业现场的安全与实时性约束，另一方面为后续与行业数据空间或可信数据空间平台对接提供了结构化接口与语义基础，可作为后续标准化和工程实践的参考蓝本。

（4）结合典型工业场景给出说明性工程案例，展示框架的落地路径与适用边界。

本文选取装配与人机协作等典型工业场景，从业务流程—数据链路—系统组件三个层面，对工业具身智能操作数据在采集、处理和上送数据空间过程中的关键路径进行抽象建模。通过案例分析，说明所提框架如何在工业现场组织数据采集、处理、资产化与可信共享流程，并为后续试点部署和定量评估提供结构化参考。

21 工业具身智能操作数据问题建模

21.1 工业具身智能操作数据的定义与边界

在本文中，我们将“工业具身智能操作数据”（以下简称“操作数据”）定义为：围绕具身智能体在工业场景中执行某一次完整的操作过程（episode）而采集的多源时序数据集。操作数据是将具身智能中的“感知—决策—执行”闭环固化为可计算对象的载体，该集合至少包含可重构的动作轨迹，与这些轨迹在时间和空间上对齐的多模态观测流，刻画任务语义、工艺参数与质量结果的高层标签三类核心要素。

与传统工业数据相比，操作数据在关注层级和结构上具有明显差异。传统 MES/SCADA 链路更侧重设备监控、过程执行、报警处置、工单流转和报表汇总，其数据通常以点位、事件、工序记录和业务单据为主；而操作数据位于‘低层控制信号’与‘高层业务数据’之间，要求在同一 episode 内，将机器人本体状态、多模态传感、工艺标签与质量结果以统一主键和时间基准组织起来，从而既能支撑学习与评估，又能与企业现有的质量管理与工艺体系对接。换言之，MES/SCADA 回答的是“生产过程是否按计划运行”，而工业具身操作数据进一步回答“机器人在一次具体操作中看到了什么、做了什么、受到了什么接触反馈，以及结果为何成功或失败”。

与数字孪生仿真数据相比，操作数据与之既存在交集，也存在本质区别：一方面，数字孪生强调通过持续采集实时数据，在虚拟空间中构建与物理资产或生产系统一一对应的“实时数字映射”，支持预测性维护、排程优化和“what-if”分析等应用；另一方面，具身智能领域的仿真数据虽可看作数字孪生的一种特例，但往往采用理想传感与控制假设，缺乏真实设备的噪声特性、工艺扰动和质量反馈。

本文所讨论的操作数据可以同时包括来自真实工厂的 episode 数据和来自高保真仿真的 episode 数据，但它们必须满足一个共同特征：围绕具体操作 episode 组织，具备可重构的动作轨迹、多模态观测以及与工艺、质量和设备状态相连接的语义标签。

2.1.2 四维操作数据分析框架

本研究将工业具身智能操作数据抽象为一个由四个正交维度构成的分析框架，分别是：数据类型维度、采集架构维度、处理流程维度和共享机制维度，如图 2 所示。

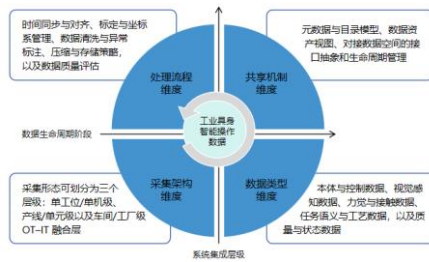


图 2 工业具身智能操作数据的四维分析框架

Fig.2 Four-dimensional analysis framework for industrial embodied intelligence operational data

在数据类型维度上，我们将操作数据拆解为五类核心要素：本体与控制数据、视觉感知数据、力觉与接触数据、任务语义与工艺数据，以及质量与状态数据。本体与控制数据主要包括关节角、速度、力矩/电流、末端位姿、控制模式和轨迹指令；视觉感知数据则覆盖 RGB、深度、点云和多视角观测；力觉与接触数据通过六维力/力矩、触觉阵列、振动和声学信号刻画接触过程中的细微变化；任务语义与工艺数据用于描述“机器人在执行什么任务、处于哪一道工序和工步、采用何种工艺配方”；质量与状态数据则通过合格判定、不良类型、返工记录及设备、班组、班次状态等信息，将操作 episode 嵌入企业质量管理和生产管理体系之中。五类要素共同决定了操作数据在“可训练、可评估、可追溯”三个层面的上限。

采集架构维度关注的是“数据从哪里来、如何进入数据平台”。从空间和系统结构上看，可将采集形态划分为三个层级：单工位/单机级、产线/单元级以及车间/工厂级操作技术（Operational Technology, OT）-信息技术（Information Technology, IT）融合层。在单工位级，机器人控制柜、PLC、外部相机、力传感器、工装夹具和本地人机界面（Human-Machine Interface, HMI）等构成最基本的采集单元，需要通过现场总线、工业以太网或采集盒子，从多源设备中尽可能低侵入地获取原始信号；在产线/单元级，多个工位和机器人通过单元控制器或产线边缘服务器汇聚数据，需要解决跨工位的时间同步、标识对齐与拓扑管理问题；在车间/工厂级，操作数据必须与 MES、SCADA、工业互联网平台以及企业数据湖连接，实现从“设备信号级”到“业务过程级”的映射，同时遵守 OT 区域、隔离区（Demilitarized Zone, DMZ）和 IT 区域的安全边界。

处理流程维度回答的是“如何把杂乱的现场信号变成结构化、可用的操作样本”。该维度可以概括为五个连续环节：时间同步与对齐、标定与坐标系管理、数据清洗与异常标注、压缩与存储策略，以及数据质量评估。时间同步与对齐需要在软时钟与硬件时钟之间做出权衡，将不同频率、不同延迟的数据对齐到统一时间轴；其中，精确时间协议（Precision Time Protocol, PTP）/网络时间协议（Network Time Protocol, NTP）等网络授时机制更适合用于工位级、产线级或跨设备的统一时间基准构建，而对于高频力觉、触觉、相机曝光等对相位一致性敏感的模式，还应结合外部触发线、同步脉冲、共享采样时钟、触发输入 / 输出

(Input/Output, I/O)或带硬件时间戳的采集卡/相机实现硬件级同步,并在必要时通过事件锚点和事后校正算法进一步完成精对齐;标定与坐标系管理通过保持多坐标系之间的标定关系和版本信息,使操作数据能够在三维空间上被准确解释和复现;数据清洗与异常标注则针对传感器漂移、通信丢包、动作中断、人工干预和设备故障等情况进行识别和标记,避免“脏数据”污染下游训练;压缩与存储策略在高频信号与有限存储之间寻求平衡,通过降采样、关键帧抽取、增量编码等方法控制数据体量,并规划在线、近线和离线分层存储;其中,对于高频本体状态、控制轨迹和结构化工艺参数,宜采用基于 Apache Parquet(一种列式存储格式, Parquet)/Apache Arrow(一种跨语言的内存数据格式, Arrow)的分块列式存储或企业数据湖表格式统一落盘,并按日期、工位/产线、任务类型、质量等级等维度进行分区组织,避免按单个 episode 拆分为大量零散 JavaScript 对象表示法(JavaScript Object Notation, JSON)文件;视觉、点云、音频等大体量模态则更适合以对象存储或文件系统形式保存,并通过引用字段与 episode 主记录关联。最后,数据质量评估通过完整性、一致性、噪声水平和标定可信度等指标,为每个 episode 或信号段打上质量标签,决定其能否进入训练库、评估库或仅作为审计和追溯用途。

共享机制维度则从“如何对外暴露和共享这些数据”的角度出发,为后续接入可信数据空间奠定基础。首先,元数据与目录模型需要在工厂内部对操作数据进行系统性编目。其次,数据资产视图将 episode、批次或任务组封装为对外可管理的“数据资产”,在字段裁剪和脱敏的基础上,为内部研发、供应链协同和第三方服务分别提供不同的视图配置。最后,对接数据空间的接口抽象和生命周期管理,则要求在不预设具体标准的前提下,设计出基于 Connector/应用程序编程接口(Application Programming Interface, API)/Adapter 的统一抽象:一端接收已处理和分级的操作数据集与资产视图,另一端以标准化元数据和访问端点的形式暴露给数据空间,同时在数据获取、使用、归档和下线的全过程中记录审计日志和使用记录,使每一份操作数据资产的流通路径和使用历史可被追踪。

综上,四维操作数据分析框架将“具身智能研究视角下的数据要素”和“工业工程视角下的系统约束与治理需求”有机统一起来:数据类型维度定义“需要采什么”;采集架构维度界定“从哪里采、如何采”;处理流程维度回答“如何把原始信号加工成标准操作样本”;共享机制维度则面向“如何在企业内部与跨组织场景中安全、可控地共享这些样本”。

21.3 操作 episode 要素模型与示例 schema

在前文的定义与分析的基础上,操作数据在逻辑上可形成围绕操作 episode 的一组关联子表。一个 episode 对应一次具有明确起止边界的操作过程。围绕这 episode,可以划分出若干功能上相对独立、结构上相互关联的子表,各子表通过统一的 episode_id 以及 time、robot_id、process_id 等主键与外键进行关联。

形式化地看,可以将 episode 元数据视作“主表”,为每次操作赋予唯一的 episode_id,并记录其起止时间、所属工单/批次、关键工序等基础信息;机器人轨迹子表以(episode_id, time)作为复合主键,存放各时间点的机器人关节角、末端位姿和控制模式等;视觉感知子表则以(episode_id, time, camera_id)作为主键,关联到具体的相机配置;力觉与接触子表可采用(episode_id, time, sensor_id)形式的键值,保证后续可以按时间与空间对齐;任务与工艺子表通过 process_id、step_id 等外键与企业 MES/工艺管理系统中的工艺路线对接;质量与运行状态子表则通过 episode_id 与工单/检验记录中的 quality_record_id 关联,为每次操作提供来自质量体系的最终判定和不良描述。这样的子表设计一方面保持了时序数据的高分辨率与多模态特性,另一方面又通过主键/外键的约束,将操作数据嵌入到既有业务数据模型之中。

在实际工程实现中,为便于跨场景复用和与开放数据集对齐,本文将上述结构进一步抽象为一个简化的 episode schema。表 1 给出了一个面向工业具身智能的示例 schema,包含约

十余个核心字段，后文对公开数据集的对齐分析和参考架构设计均以此为“标准字段集合”。

表 1 工业具身操作 episode 示例 schema

Table 1 Example schema of an industrial embodied operation episode		
字段名	类型示例	说明
episode_id	字符串类型 (String, STRING)	操作 episode 全局唯一标识，用作各子表主键或外键的核心关联字段
workorder_id	STRING	所属工单或订单编号，用于将操作数据映射到生产任务与批次
process_id	STRING	工序标识，对应工艺路线中的具体工序（如装配、拧紧、检测等）
step_id	STRING	工步或子任务标识，用于 finer-grained 的任务分解与对齐
robot_id	STRING	执行该 episode 的机器人或具身智能单元标识
start_time	时间戳类型 (Timestamp, TIMESTAMP)	episode 起始时间（统一时间基准，用于对齐多模态数据）
end_time	TIMESTAMP	episode 结束时间，可与 start_time 共同估算持续时长
traj_series	数组类型 (Array, ARRAY)<结构体类型 (Struct, STRUCT)>（逻辑上可抽象为 ARRAY<STRUCT>） / Parquet 或 Arrow 分块存储（物理推荐）	压缩后的本体与控制轨迹序列（包含 time、joint_pos、tcp_pose、control_mode 等字段）；逻辑上可抽象为 ARRAY<STRUCT>，物理层面宜采用 Parquet/Arrow 等分块列式格式统一存储，以支持批量扫描、并行加载和训练阶段高吞吐读取，而不宜按 episode 直接序列化为零散 JSON 文件。
vision_series_ref	STRING/统一资源标识符 (Uniform Resource Identifier, URI)	指向视觉/点云数据序列的引用（如文件路径、对象存储键或流 ID），本身不内嵌大体量图像数据
force_series_ref	STRING/URI	指向力/力矩等接触数据序列的引用，用于接触过程重建与分析
task_semantic	STRING/JSON	任务与工艺语义标签，可包含任务名称、自然语言描述、作业指导书段落引用、阶段/子任务状态、关键事件标记等，用于刻画长序列操作中的语义转移与意图变化
process_params	STRUCT/映射类型 (Map, MAP)（逻辑表示） / Parquet 列字段、Map/Struct 列或参数子表（物理推荐）	与本次操作相关的关键工艺参数及上下文状态摘要（如扭矩设定值、速度/加速度、工艺配方版本、工装状态、来料批次、环境约束等），用于对影响策略决策的部分隐含变量进行显式登记；逻辑上可表示为结构化键值对或结构体，物理层面宜优先落为 Parquet 列字段、Map/Struct 列或独立参数子表，便于检索、过滤与批量训练加载；JSON 仅建议用于接口交换、轻量导出或调试视图，不宜作为训练主存储格式。
quality_label	STRING	质量结果标签，如“OK”“NG”，可按企业质量体系统一编码

字段名	类型示例	说明
defect_type	STRING	不良类型或原因代码，可为空，用于后续质量分析与模型训练
equipment_state	STRING/JSON	episode 期间的关键设备状态摘要（如运行/故障/待机、异常码等），必要时可链接到更细粒度日志
operator_info	STRING/JSON	班次、班组、操作员标识等人因信息，视隐私与合规要求决定是否保留或脱敏
data_quality_level	STRING/枚举类型 (Enumeration, ENUM)	数据质量分级结果，如‘A/B/C’等，用于区分主训练库、评估库、异常样本库/困难样本库与审计库的可用范围；其中异常或失败样本可在原因清晰、标注完备和边界可控的前提下，进入面向鲁棒性增强、异常检测和恢复学习的专用数据子集

表 1 所示 schema 并非要求在每个场景中所有字段都为必填，而是为工业具身操作数据提供一个“收敛的目标结构”。在最小配置下，episode_id、robot_id、start_time、end_time、traj_series、task_semantic 和 quality_label 等字段构成了具备基本可用性的 episode；在条件允许时，可以逐步补充 vision_series_ref、force_series_ref、process_params、defect_type、equipment_state 等增强字段，以提升数据在策略学习、工艺优化和质量分析中的价值。需要进一步说明的是，表 1 给出的主要是逻辑字段集合，而非要求所有字段采用同一种物理落盘方式。对于 traj_series、force 等高频结构化时序数据，宜采用 Parquet/Arrow 等可分块、可分区的列式格式统一存储；对于 process_params 等结构化工艺字段，宜优先采用列字段或参数子表方式组织；对于图像、点云、视频、音频等大体量模态，则继续采用对象存储或文件系统保存，并通过 URI/ref 字段与 episode 主记录关联。对于长时序、接触丰富且存在明显阶段切换的工业操作，仅以静态 process_params 描述上下文仍然不够；在该 schema 基础上，还可进一步扩展阶段/子任务标签、事件锚点、意图状态以及环境上下文摘要等字段，并将标准作业程序（Standard Operating Procedure, SOP）、工艺卡或作业指导书中的步骤约束映射到 episode 的阶段边界与事件序列中。这样既可避免形成海量小文件和 JSON 碎片，也更有利于数据湖管理、批量训练加载和跨系统共享，并为后续利用大语言模型从非结构化作业文本中抽取任务逻辑并形成机器可解释的时序约束提供接口。

3.2 代表性具身操作数据集的对比分析

本章选取 RLBench、DROID、RH20T 与 CALVIN 四个在具身操作和机器人操作学习领域具有代表性的数据集，作为对比分析的对象。

3.2.1 基于四维框架的数据类型与 episode 结构对比

本节首先从“数据类型维度”出发，对四个数据集在本体与控制、视觉感知、力觉与接触、任务语义与工艺、质量与状态五类要素上的覆盖情况进行横向对比；同时，对比其 episode 组织方式与显式标注程度，为后文提炼“可迁移模式”和“工业缺口”奠定基础。

本体与控制数据方面，四个数据集都提供了相对完整的机器人状态与控制信息。RLBench 在仿真环境中输出关节角、关节速度、末端位姿、抓取器开合状态等本体数据，配合任务级动作命令，便于在同一仿真环境中精确重放演示轨迹，实现“可重构”的 episode 结构。DROID 在真实 Panda 平台上记录了示教轨迹和机器人动作，覆盖多机构、多场景下的轨迹演示，支持在统一硬件模板上重现交互过程。RH20T 为每条序列记录动作信息和本

体状态，强调在真实接触交互下的轨迹变化。CALVIN 则在仿真环境中提供与动作序列对齐的状态表示，用于长时序策略学习。从可重放性和时间结构清晰度看，RLBench 与 CALVIN 的 episode 边界和步骤划分最为规则，而 DROID、RH20T 的 episode 更接近真实操作记录，长度和结构具有显著变化。

在视觉感知方面，四个数据集均提供了丰富的图像模态，但覆盖方式有所差异。RLBench 通过多视角 RGB/Depth/Segmentation 等观测为每个任务配置若干虚拟相机，且各视角严格时间对齐，是典型的“理想传感”配置。DROID 在真实场景中普遍采用三台 ZED 相机等多视角相机阵列，记录高分辨率 RGB 流，并附带标定信息，在跨环境多场景下提供统一视觉观测接口。RH20T 为每条操作序列采集多视角视觉数据，覆盖不同机器人、视角和背景，显著强调视觉多样性。CALVIN 则采用单机载相机视角，结合仿真中可控的环境变化，侧重于“视觉-语言-动作”三者在长时序上的一致性。

力觉与接触数据是区分各数据集的重要维度。RLBench 与 CALVIN 主要工作在仿真环境，并未单独提供真实世界的六维力/力矩或触觉阵列等通道，接触信息更多隐含在视觉与状态变化中。DROID 的核心关注在于大规模、多环境的视觉与本体轨迹，目前公开描述中主要强调视觉与关节/控制模态，对外置力/触觉通道着墨较少。相比之下，RH20T 明确将“接触丰富、多模态”作为设计目标，每个 episode 同时包含视觉、力觉、音频和动作信息，并强调通过多视角相机和校准机制精确捕捉接触过程，是四者中在“力觉/接触事件刻画”最为完备的数据集。

在任务语义与语言层面，四个数据集也呈现出不同的优势。RLBench 为每个任务分配任务 ID 和变体 ID，提供任务级和阶段级标注，并在部分扩展版本中配有可读的任务名称和描述。DROID 为轨迹标注了任务类别、场景 ID 等元数据，并在部分采集流程中引入自然语言指令或说明，支持“语言条件操作”的研究。RH20T 在每条机器人序列之上配套一条人类演示视频和对应的语言描述，形成“机器人操作—人类示范—语言讲解”三元组，在任务语义丰富度和跨模态对齐方面具有独特优势。CALVIN 则以长时序语言条件任务为核心，将自然语言指令序列、视觉观测与复杂动作序列强绑定，是四者中对“语言条件策略”和“多轮指令—子任务链路”建模最系统的基准。

质量结果与设备状态是与工业需求最直接相关、但在现有开放数据集中普遍较弱的一环。四个数据集都以“可用演示”为设计目标，episode 多为成功执行的示例，质量结果往往以“任务成功/失败”的隐含条件存在：RLBench 与 CALVIN 的示教与评测环境通常默认使用成功轨迹，缺乏显式的不合格类型或返工信息；DROID 在采集中对明显失败的演示进行过滤或截断，公开版本聚焦于高质量演示轨迹，同样很少提供面向质量管理的细粒度标签；RH20T 虽然是接触丰富数据集，但质量语义主要体现在任务完成与否及语言描述中，尚未系统映射到类似“合格/不合格”“不良类型”“返工路径”这类工业质量体系概念。这意味着从工业 episode schema 的视角看，四个数据集在“质量与状态”这一列整体偏空，需要在工业应用中增补与工艺路线、质量检验、设备运行状态相结合的标签。

表 2 对四个数据集在数据类型与 episode 结构层面的覆盖情况进行了汇总。可以看到：RLBench 与 CALVIN 在 episode 结构清晰性、本体与视觉协同、任务/语言语义建模方面较为突出；DROID 在真实场景多样性和跨机构协同采集方面具有优势；而 RH20T 则在力觉与接触事件、多模态同步和“机器人-人-语言”三元对齐方面提供了最丰富的要素。然而，从工业具身智能的角度看，四者普遍缺乏与工序号、工艺参数、质量判定、设备状态及生命周期信息深度绑定的设计。

表 2 代表性具身操作数据集在数据类型与 episode 要素上的覆盖对比

Table 2 Comparison of representative embodied manipulation datasets in terms of data types and episode element coverage

数据集	本体与控制数据	视觉感知数据	力觉/接触数据	任务语义/语言	质量与状态要素	episode 结构与失败样本概况
RLBench	仿真关节状态、末端位姿、抓取多视角器状态与控制命令, 支持精确重放	RGB/Depth/Segmentation, 严格对齐	无显式真实力觉/触觉通道	任务 ID、变体 ID、默认阶段标注, 部分任务有名称/描述	聚焦高质量演示, 质量结果多隐含在任务完成条件中	episode 边界清晰, 演示轨迹与离散步骤结构化良好, 基本不提供系统化失败样本
DROID	真实机器人示教轨迹, 多机构多环境统一模板	三相机等多视角 RGB(部分含深度), 带标定信息	以本体与控制为主, 无专门外力/触觉通道公开说明	任务类别、场景 ID 等元数据, 部分采集含语言指令	关注任务完成与否, 缺乏面向工业质量管理	episode 为真实示教轨迹, 长度可变, 多场景“in-the-wild”, 失败或中断样本一般被过滤或未细粒度标注
RH20T	真实机器人动作与本体状态, 多机器人、多技能覆盖	多视角视觉序列, 涵盖多种场景与构型	明确提供力觉与音频等接触相关信息, 是接触丰富操作的代表	每条机器人序列对应一条人类演示视频和语言描述, 语义信息丰富	关注任务完成与否, 缺乏面向工业质量管理的显式不良类型、返工路径等标签	episode 明确、数量大, 强调多模态接触过程, 但无系统化工工业质量/设备状态标注, 失败样本多以“未完成”隐式呈现
CALVIN	仿真状态与动作序列, 支持长时序语言条件策略学习	机载视觉观测, 环境变化可控	无独立真实力觉通道, 信息隐含于状态变化	多轮自然语言指令与子任务序列紧密对齐, 是语言条件建模最强的一类	主要用于评测任务成功率, 无工业意义上的质量/状态标签	episode 边界与子任务分解清晰, 长时序结构显式, 失败多体现在评测指标中而非 episode 内部质量字段

批注 [1]:

通过对比可以看到, 如果仅从“具身智能算法研究”的角度出发, 四个数据集已经覆盖了相当丰富的本体、视觉、语言和部分接触要素, 然而它们在质量与状态要素、与工艺/工序

的对接以及及设备/班组/班次信息的耦合方面仍存在系统性缺口。

3.2.2 采集架构与处理流程实践

采集架构与处理流程方面，RLBench、CALVIN、RH20T 与 DROID 可分为三类典型模式：仿真环境中的“理想同步”采集，实验室平台中的高保真多传感器采集，以及 in-the-wild 条件下的大规模分布式采集。对于 RLBench 和 CALVIN 这类仿真基准，其采集架构的核心特点是“逻辑上集中、物理上无噪”的理想化假设。RLBench 在 CoppeliaSim 等仿真引擎内统一生成本体状态、多视角 RGB/Depth/Segmentation 等观测信息，所有传感与控制数据共享同一个仿真时间基准和更新频率，天然不存在网络抖动、时钟漂移或不同采集进程之间的异步问题；平台在数据导出时直接以任务为单位打包 episode，将演示轨迹、多模态观测和任务标注结构化为统一格式。CALVIN 作为长时序语言条件任务的仿真基准，同样在统一仿真环境中生成视觉观测、动作序列和语言指令，采集流程基本等价于“从单一仿真主循环中顺序截取时间片”，各模态之间的时间对齐几乎是“零成本”的。这种同步架构极大地便利了算法开发与评测，但与真实工厂中“机器人控制器+PLC+外部相机+力传感器+MES/SCADA”多源异构、分布式网络架构之间存在明显鸿沟：在 OT-IT 分区、现场总线与工业以太网共存、设备各自维护本地时钟的前提下，时间同步与坐标对齐需要专门的系统设计和元数据记录，不能依赖仿真中的隐式假设。

相较于仿真环境，RH20T 所采用的是典型的“实验室级高保真多传感器采集架构”。在采集流程上，RH20T 为每条操作序列记录视觉、力觉、音频和动作等多模态信号，并特别强调研究团队投入了大量精力进行多相机标定与坐标系对齐，确保每个 episode 中不同视角的观测在空间和时间上都能够复现真实接触过程。从工业视角看，要在真实物理环境中构建高质量操作 episode，仅仅“接上传感器”远远不够，还必须将传感器布局 (in-hand vs global)、多视角标定链路、力觉与视觉的时间同步等基础工作流程化、制度化；对于高频接触过程，宜优先通过硬件触发线、同步采样卡或共享时钟实现相机与力传感器的同步采集，并将触发方式、同步误差范围、标定版本与生效时间记录为元数据的一部分。实验室平台通常假定所有设备在同一物理空间、由同一团队维护，网络拓扑简单；但工业车间则必须在 OT 安全域、分段网络和既有 IT 系统约束下，实现类似的多传感器集成，这对采集架构提出了更高的工程要求。

DROID 所采用的则是大规模 in-the-wild 分布式采集模式，其重点更多在“多场景部署的一致性”和“采集流程的可复制性”上。在采集架构上，每个现场通过标准化硬件模板和图形用户界面 (Graphical User Interface, GUI) 完成演示记录、元数据标注和成功/失败标记；在处理流程上，则大量利用 ZED 相机从压缩 MPEG-4 第 14 部分媒体文件格式 (MPEG-4 Part 14, MP4) 视频中恢复深度信息的能力，从而在保证视觉-深度对齐前提下降低存储和传输压力。尽管 DROID 并不涉及 MES/SCADA 等企业系统，但其“轻量采集包+标准操作流程+分布式异步上传”的模式，对工业场景下在多产线、多工厂部署统一数据采集网关具有重要启示：一方面，通过标准化硬件/软件栈和操作规范，可以在不依赖单一中心团队的情况下快速扩展采集覆盖范围；另一方面，通过在本地完成必要的预处理和压缩，再以统一格式异步汇聚到中央数据平台，可以在不改造现有 OT-IT 结构的前提下实现大规模数据汇聚。

上述三类模式对工业具身操作数据采集与处理具有直接指导意义：采集架构与处理流程不应仅被视为“数据前端”，而应作为操作数据体系设计的核心部分，从一开始就围绕 episode 要素模型、质量分级和后续共享/治理要求来规划。第一，仿真数据集所体现的‘统一时间基准+清晰 episode 边界+一致的数据组织结构’应当成为工业采集架构的目标特性之一，但在现实 OT-IT 环境中，这样的理想状态不能仅依赖软件层面的统一授时，而应通过‘统一时基+硬件触发同步+标定版本管理+严格元数据记录’来工程化重建：其中，PTP/NTP 等机

制主要用于跨设备粗同步和系统级统一时基，高频视觉、力觉、触觉等模态则宜通过外部触发线、同步脉冲或带硬件时间戳的采集设备完成硬件级对齐，并辅以事后误差校正。第二，RH20T 展示了在单个平台上进行多视角、多模态高保真采集所需的传感器布局与标定流程，这为工业车间设计“单工位/单元级具身数据采集节点”提供了模板。第三，DROID 的分布式 in-the-wild 采集经验表明，要在多场景、多组织中持续收集高质量 episode，通过标准化采集套件和低门槛操作界面降低现场负担，通过统一的元数据规范和轻量处理流程保证汇聚到中心的数据在结构和质量上的可控性。

32.3 共享与复用机制分析

从共享与复用的角度看，RLBench、DROID、RH20T 和 CALVIN 这类公开数据集，已经在“如何发布、如何访问、如何被社区长期复用”方面形成了一套事实上的通行做法。本节重点分析开放数据集与工业数据空间之间的关键差距，并总结出当前工业领域中存在的空白与其可借鉴的共享与复用模式。

从发布与许可机制来看，四个数据集均采用了“公开访问+明确许可+标准引用”的组合模式。RLBench 的代码与环境在 GitHub 上以专用许可发布，同时社区中已有基于 RLBench 预生成演示数据的镜像数据集，以阿帕奇 2.0 开源许可协议 (Apache License 2.0, Apache-2.0) 等开源许可证形式在 Hugging Face 上提供，方便研究者直接下载并复现实验流程。DROID 在其官网和论文中明确声明完整数据集以知识共享署名 4.0 国际许可协议 (Creative Commons Attribution 4.0 International License, CC-BY4.0) 授权发布，并提供交互式可视化工具、策略训练代码、预训练策略以及硬件搭建指南，便于他人基于该数据集进一步研究和扩展。RH20T 将数据划分为子集，其中 RH20T-NC 以 CCBY-NC4.0 发布，允许非商业免费使用但禁止商业用途，并在官网进一步说明了重分发条件和咨询方式。CALVIN 则以 MIT 许可发布代码和数据，配套清晰的引用格式和基准说明，降低研究者使用门槛。这种“开放许可证+明确引用规范+统一入口网站/仓库”的组合，事实上构成了研究社区中具身操作数据共享的基础规则。

在复用模式与生态建设方面，这些数据集同样展现出较强的“平台化”特征。RLBench 通过任务定义、数据生成脚本和预生成数据，成为 PerAct、RVT、GENIMA 等大量工作复现实验的标准基准，不同论文通常共享同一套任务划分与数据切分方式。DROID 提供了强化学习数据集规范 (Reinforcement Learning Dataset Specification, RLDS)/原始数据两种格式，并配套专门的策略学习仓库，使其不仅是一个静态数据集，而是围绕硬件模板、控制栈、数据组织和训练流程的完整解决方案。RH20T 通过“机器人轨迹+人类演示视频+语言描述”的三元结构，成为研究 one-shot imitation 和多模态对齐的核心基准，相关论文和开源实现通常在其数据组织方式上进行扩展。CALVIN 则围绕语言条件长时序任务，形成了 HULC 等一批方法的共用评测平台，在任务拆分、指令设计和环境变体组织上形成了事实标准。从复用形态看，这些开放数据集强调的是“可复现+可扩展”：通过统一任务定义、基准指标和参考实现，让后续算法工作可以在同一数据组织和接口层上迭代。

然而，从“数据目录/数据资产”视角进行分析，开放数据集与工业数据空间之间仍然存在若干结构性差距。首先，这类数据集的“目录”和“元数据”主要服务于研究者下载和训练，这些做法已经具备“内部数据目录”的雏形，但与工业数据空间对目录的要求仍有明显差异：工业场景需要按工序、工艺路线、设备资产、生产批次等维度进行检索，并要求目录项中至少包含质量分级、有效期、责任主体和适用用途等属性，形成可被审计和追责的“数据资产登记簿”，而这些在公开数据集中基本空缺。

其次，从“数据资产视图和访问控制”的角度看，开放数据集更多在于“统一公开视图”，

而非面向多角色的差异化视图管理。对于 RLBench、DROID、RH20T 与 CALVIN，数据发布通常是“一视同仁”的：任何满足许可条件的用户均可在相同粒度上下载全部或子集数据，访问控制主要体现在“是否允许商业使用”“是否需要注明来源”等许可条款上。工业数据空间则需要将 episode/批次/任务组抽象为不同粒度的数据资产，针对内部研发、供应链协同、第三方服务等不同角色，提供字段裁剪、脱敏策略和使用限制不同的视图，并通过数字合约或策略引擎在访问接口层强制执行。这种“多视图、多角色、策略驱动”的资产管理逻辑，在现有开放基准中鲜有体现。

第三，在访问审计与生命周期管理方面，开放数据集基本仅要求在学术规范下完成引用和鸣谢，生命周期主要体现为版本更新和镜像维护。而工业数据空间需要在‘在线运行→近线分析→离线归档→选集入空间’的生命周期各阶段上，对数据的可见范围、质量要求和保留策略进行精细管理，并在访问过程中留下可追溯的审计记录，支撑合规检查、责任追踪和价值结算。进一步看，在工业企业以操作数据换取算法能力或优化服务的场景中，风险并不止于“是否越权访问”，还包括“是否通过合法访问形成了超出约定范围的模型能力和收益”。因此，工业具身操作数据进入数据空间后，不仅需要明确所有权、使用权和责任边界，还需要在合约与治理机制中体现收益分配、服务边界、衍生模型复用边界以及对工艺敏感信息的保护要求，避免第三方通过大量 episode 学习反向推断企业关键工艺参数、质量阈值或设备能力边界，造成事实上的技术失密。

上述分析可归纳为表 3 所示的**启示与空白**：一列是开放基准实践中可直接迁移到工业具身操作数据建设中的模式，另一列是工业数据资产体系中必须补齐的关键空白。

表 3 开放具身数据集共享机制对工业数据资产设计的启示与空白

asset design		
维度	可直接借鉴的开放基准模式	工业数据资产必须补齐的空白
数据组织与 episode 抽象	以任务为单位组织 episode，统一命名规则和目录结构，提供标准化下载与加载接口（如 RLDS、预生成 NumPy 压缩数组格式 (NumPy Zip, npz)/分层数据格式第 5 版 (Hierarchical Data Format 5, HDF5/hdf5)）	在 episode 层引入工单、工序、批次、设备等业务主键，将操作 episode 显式嵌入企业生产数据模型，并与质量记录和设备维护记录关联
		构建面向工艺与业务的“操作数据目录”，包含工艺参数、质量分级、责任主体、适用用途和有效期等字段，支撑资产登记与数据空间目录对接
元数据与目录	型等元数据支持检索与过滤	在企业内部构建“数据产品”视角，将操作数据资产与分析应用、具身策略服务打包为可复用服务单元，引入服务等级协议 (Service Level Agreement, SLA)、版本管理和跨项目复用机制
复用与生态	“数据集+代码+基准指标+参考实现”一体化发布，形成算法社区的事实标准和复用生态（如 RLBench/CALVIN 上的大量后续工作）	在技术层面落地细粒度访问控制与使用控制：按角色和用途区分视图
许可与访问控制	通过 MIT、Apache-2.0、CC-BY/CC-BY-NC 等统一开源或开放数据许可，简单区分“是否可用”“是否需署名”	策略引擎或数字合约控制字段可见性、使用频次和用途范围
审计与生数据一般视为“静态开放资源”，版本更新通过建立“在线-近线-离线-数据空间选		

维度	可直接借鉴的开放基准模式	工业数据资产必须补齐的空白
命周期	仓库和官网公告，审计主要依赖学术引用规范集”的生命周期管理机制，在访问与流通过程中记录审计日志、数据质量演变和价值创造轨迹，支撑合规监管与价值结算	

从本节的对比可以看出，开放具身数据集在“如何发布并让研究社区长期复用一份操作数据”这一问题上，已经给出了成熟的答案：围绕清晰的 episode 抽象、统一的目录与加载接口、开放许可和配套代码，形成可复现、可扩展的开放基准生态。而工业数据空间需要解决的问题，则是如何在此基础上叠加工艺语义、质量分级、资产边界、访问审计和合规治理等维度，将原本“为算法研究而整理的数据”提升为“可被管理、可被追责、可被结算的生产要素”。

4.3 工业具身操作数据采集与处理参考架构

4.3.1 总体架构与分层

在前文四维操作数据分析框架和代表性数据集对比的基础上，可以将工业具身操作数据从采集到资产化的全过程，抽象为一个自下而上的分层参考架构。整体上，该架构由三层构成：底层是面向具体设备和工位的采集与边缘处理节点，中间层是面向车间/工厂的数据汇聚与处理层，上层则是围绕操作 episode 构建的数据资产库与内部数据目录。三者协同工作，将原本分散在机器人控制器、传感器、PLC 和业务系统中的碎片化信号，逐步加工为可训练、可评估、可共享的操作数据资产。

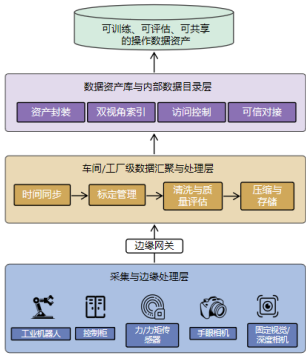


图 3 工业具身操作数据采集与处理参考架构总览

Fig.3 Overview of the reference architecture for industrial embodied operation data acquisition and processing

在底层，单工位/单元采集节点是整个架构的“前端触点”，直接面向具身智能体与物理环境。典型构成包括：一台或多台工业机器人及其控制柜，安装在手爪或末端的力/力矩传感器和手眼相机，覆盖操作区域的固定视觉/深度相机，必要的触觉、振动、声学传感器，以及负责与现场 PLC、I/O 模块交互的接口单元。这些设备通过现场总线、工业以太网等方式接入本地边缘网关（edge gateway），由边缘网关统一完成数据采集、时间戳标记和初步预处理。与 RLBench、CALVIN 这类理想仿真环境不同，工业现场必须假设各设备存在独立时钟和不同采样频率，因此边缘节点需要负责软/硬件时钟同步、简单的异常过滤和必要的

脱敏操作，将多源信号整理为围绕 episode_id 的“原始但结构化”的边缘数据包，同时保留与工序、工单和设备 ID 的映射关系。

中间层的车间/工厂级数据汇聚与处理层，是将多工位、多单元的边缘数据整合为统一操作数据视图的关键。该层通常部署在车间级边缘服务器或工厂数据中心，处于 OT 区域和 IT 区域的交界处，通过安全网关与各单元边缘节点、MES/SCADA 系统和企业数据平台进行通信。一方面，该层要基于统一时间基准（如 PTP/NTP）对来自不同工位、不同机器人和不同传感器的数据进行精确对齐，将本体与控制轨迹、多视角视觉流、力觉与接触信号等关联到相同 episode 时间轴；另一方面，还要维护标定与坐标系管理服务，对机器人基座、工具坐标系、工装夹具、相机外参等信息进行版本化管理，并在数据记录中显式写入相应的标定版本号和生效时间。在此基础上，数据清洗与异常标注模块负责识别传感器失效、通信丢包、动作中断、人工干预和设备故障等情况，对 episode 进行质量打标和分级；压缩与存储模块则根据在线、近线和离线不同层级的存储策略，对高频信号进行降采样、关键帧提取或增量编码，在保证重构能力的前提下控制数据体量。经过这一层处理之后，原始的边缘数据被转化为符合第 2 章 episode schema 的“标准操作样本”，并带有数据质量等级、异常标签和标定信息等元数据。

上层的操作数据资产库与内部数据目录层，将标准化的操作样本进一步封装为可管理、可检索、可授权使用的“数据资产”。在资产库中，每个 episode、批次或任务组都可以作为独立的数据资产单元，拥有唯一的资产 ID，并与其对应的 episode_id、workorder_id、process_id 等字段建立一一映射。内部数据目录则从“业务+技术”双重视角为这些资产建立索引：既记录数据源设备、采集时间范围、模态构成和质量等级，又记录对应工序、工艺参数、质量结果、不良类型以及责任部门等信息，使研发、工艺、质量和运维等不同角色可以按任务类型、工艺路线或质量问题等多维条件检索所需的操作数据。在访问控制上，该层还需要区分内部研发视图、供应链协同视图和第三方服务视图等不同“视图配置”，为后续与可信数据空间的连接预留接口：例如，将 episode 级资产与其元数据通过统一的 API 或连接器（Connector）对外暴露，同时结合组织策略或数字合约实现对字段可见性、使用目的和调用频次的约束，并将访问行为回流至资产库中的审计记录。

通过这样一个自下而上的分层架构，可以在工程上将四维操作数据框架和开放数据集经验“落地”为一张清晰的系统总图：底层强调如何尽可能低侵入地从机器人和传感器采集结构化原始信号，中间层强调如何通过时间同步、标定管理、清洗与质量评估，把这些信号加工为 episode 级标准样本，上层则负责在 episode schema 的基础上完成资产化和目录化，为跨系统乃至跨组织的共享与治理奠定基础。

4.3.2 从 OT 设备到操作 episode 的数据路径

本节以装配和拧紧这两类典型任务为例，说明“从 OT 设备到操作 episode”的数据路径。在装配任务中，现场通常包括一台或多台工业机器人控制柜、一套末端执行器、若干外部相机、必要的力/力矩传感器以及与生产线节拍相关的 PLC。机器人控制柜通过工业以太网或现场总线向边缘网关周期性输出关节状态、末端位姿和控制模式等本体与控制数据；外部相机和手眼相机通过千兆以太网或采集卡，将时间戳对齐的视频流或图像帧推送到边缘节点；末端或基座上的力/力矩传感器通过专用放大器或现场总线接口输出高频力觉信号；PLC 则以较低频率提供节拍信号、工序开始/结束信号以及与工装、输送线状态相关的离散量。边缘节点在 OT 区域内通过多个网络接口接入上述设备，对所有采集到的信号统一打上本地时间戳，并保留原始设备时间信息，为后续时间同步和对齐提供基础。

在拧紧任务中，数据路径的结构类似，但涉及的关键 OT 设备略有不同。机器人需要与智能拧紧枪或拧紧控制器协同工作，拧紧控制器通过现场总线或以太网接口向边缘节点提供

扭矩、角度、拧紧曲线以及合格/不合格判定等过程数据，同时通过数字量或消息接口向 PLC 报告拧紧结果。外部相机可能用于识别螺栓位置或读取条码，力/力矩传感器则用于监测接触过程中的异常。边缘节点在采集本体、视觉、扭矩曲线和 PLC 信号的同时，会根据预先配置的规则，将“一次拧紧动作”的开始与结束事件识别出来：例如，当 PLC 上某一位触发“拧紧开始”，且拧紧控制器报告一次完整扭矩曲线并给出结果，再结合机器人程序中的任务标识和工件 ID，即可判定这是一个独立的 operation episode。

在边缘与车间级处理层之间，时间同步与标定是把“信号流”转化为“操作 episode”的关键环节。首先，工厂会通过 PTP/NTP 等方式为车间服务器和边缘网关提供统一时间基准，而各设备的本地时间则通过一次性偏移标定或周期性校正映射到这一基准时间轴上。车间/工厂级处理平台接收到来自多个工位的边缘数据流后，会根据统一时间基准，将机器人轨迹、多视角视觉流、力觉和 PLC 事件对齐到同一时间线，在此基础上完成 episode 的切分和重构。其次，标定管理服务负责维护机器人基座坐标系、工具坐标系、工装坐标系和各相机坐标系之间的空间关系，并为每一次标定生成版本号 and 生效时间区间；在处理过程中，平台会根据 episode 的时间范围选择对应的标定版本，将像素坐标和力觉方向等信息转换到统一的三维参考系中，从而确保后续分析和训练时对“机器人在何处、如何接触工件”的空间理解是一致的。

清洗与 episode 构建则是把上述对齐后的多模态信号加工成可训练、可分析的“标准操作样本”的最后一步。对于装配或拧紧任务，平台会根据 episode 切分规则，将机器人轨迹、视觉帧序列、力觉时间序列以及 PLC 信号聚合到一个 episode_id 下，同时引入工单号、工序号、工步号、工艺参数版本和操作员信息等业务标签。随后，数据清洗模块会检测并标记明显异常情况，例如传感器读数固定在异常常数、通信中断导致的长时间缺失、机器人急停或人工干预、拧紧曲线明显偏离正常模式等；对于这类异常 episode，可以在质量分级中降低等级，或仅标记为“失败样本”，保留用于鲁棒性训练或安全分析。完成上述步骤后，平台将 episode 映射到第 2 章定义的示例 schema 中，形成结构化的操作样本：既包含压缩后的 traj_series、vision_series_ref 和 force_series_ref，也包含 task_semantic、process_params、quality_label 和 defect_type 等语义信息，从而可以直接作为具身策略训练、性能评估、故障诊断或工艺优化的输入。

通过这一完整的数据路径，原本散落在机器人控制器、传感器和 PLC 中的 OT 信号，被逐步抽象和重组为围绕具体操作 episode 的标准数据对象，实现了从“设备级控制日志”到“面向具身智能与工艺分析的操作样本”的转化。相同的路径也适用于码垛等其他任务，只是 episode 的边界判定可能由“抓取-移动-放置”的动作序列和堆垛层级变化决定，但本质上仍然遵循“OT 信号采集→时间同步与标定→清洗与切分→映射到 episode schema”的统一流程。

4.3.3 数据质量控制与安全约束嵌入

在工业环境中，操作数据不仅要“能采集、能重放”，还必须“质量可控、风险可管”。数据质量控制方面，基于前文的数据质量评估思路，本文提出一个简化的三档质量分级方案，将每个 episode 或批次划分为 A/B/C 三个等级。A 级数据代表结构完整、标定可信、时间同步准确且无明显异常的高质量样本：本体轨迹连续、关键传感模态齐备，episode 起止事件清晰，并有明确的质量结果标签和工艺参数记录；对于涉及高频视觉、力觉或触觉耦合的 episode，还应具有可核验的同步链路说明，例如硬件触发记录、统一时基配置、同步脉冲日志或可追溯的对齐误差评估结果。这类数据适合作为训练库和评估库的主体。B 级数据在结构上基本完整，但可能存在局部缺失或轻微异常，例如个别视觉帧丢失、小段力觉信号不稳定、标注信息不够完备等；这类数据可以进入训练库，但在评估和审计时需要谨慎使用，并在元数据中显式记录“已知缺陷”。C 级数据则指存在明显缺失、时间对齐问题、标注不可

靠或执行结果异常的样本，但其中应进一步区分两类情形：一类是因采集损坏、关键字段缺失或记录自相矛盾而难以直接复用的样本，这类数据原则上不进入主训练与自动评估，而主要保留在审计库中用于手工分析、问题追踪或存在性证明；另一类是虽然执行失败、出现异常或存在人工接管，但其记录过程真实、边界清晰、异常原因可解释的样本，这类数据在补充失败阶段、异常类型、干预方式、恢复结果和可用字段范围等标注后，可进入面向鲁棒性增强、异常检测、恢复策略学习和安全强化学习的专用数据子集。通过这种分级与分流结合的规则，可以建立‘主训练库（以 A/B 为主）—评估库（以 A 为主）—异常样本库/困难样本库（包含经筛选标注的部分 B/C）—审计库（保留不可直接复用的 C）’的分层数据使用模型，让下游算法和应用明确‘能用什么数据、在什么任务中用、以何种方式用’。

对于需要利用异常样本的训练任务，企业可进一步在内部资产库中建立‘失败样本标签集’‘人工接管样本集’或‘高风险工况样本集’等专题子集，并通过元数据、使用控制策略和实验审批流程，防止低质量异常数据被误用为通用策略训练样本。

安全约束嵌入方面，在采集与处理架构中应系统性地嵌入 OT-IT 安全域和权限控制的约束。大多数工业企业采用类似 Purdue 模型的分区理念，将现场设备和控制系统划分在 OT 安全域（如 Level 1-2），而车间级边缘服务器、MES/SCADA 等位于中间域或 DMZ，企业 IT 系统和数据中心则位于更高层级。这样的安全分区意味着：第一，单工位/单元采集节点通常只能在 OT 区域内部通过只读接口获取控制器、相机和传感器数据，不能反向向控制系统发送未经授权命令；第二，OT 区域向上层发送数据的路径必须经过边界防火墙或安全网关，采用白名单协议和端口、必要的加密和身份认证，并尽量采用‘边缘预处理+异步/批量上送’的模式，减少对实时控制网络的影响；对于对时延和抖动高度敏感的任务，数据空间连接器不应进入毫秒级控制闭环，现场控制仍应由 OT 域内控制器、本地 PLC 或边缘节点承担，数据空间主要服务于数据上送、近线分析、离线训练、审计与治理。第三，车间/工厂级处理平台在访问和处理操作数据时，需要引入角色与权限模型，对不同用户群体开放不同粒度的访问能力，例如允许数据工程师访问 A/B 级训练样本的全量技术字段，但对外部顾问或第三方服务仅开放脱敏后的质量统计与有限字段视图。在操作数据架构中显式考虑这些安全约束，会直接影响组件的位置与接口设计。面向可信数据空间的连接器也应部署在 DMZ 或 IT 域，严格遵守‘只读、最小必要字段、策略可审计’的原则，避免从外部通道绕过企业自身的访问控制体系，也避免通过外部链路直接干预现场执行器或低层控制回路。

通过在质量分级和安全域划分两个维度同时“前置设计”，可以保证工业具身操作数据既具备足够的技术价值，又在采集、处理和共享的全链路上满足企业对安全、合规和责任追踪的要求，为后续在可信数据空间中的可控流通打下基础。

5.4 面向可信数据空间的共享与治理设计

5.4.1 操作数据资产化与元数据建模

在内部数据目录与 episode schema 的基础上，要让工业具身操作数据真正进入可信数据空间，第一步是完成从“内部技术样本”到“对外可见数据资产”的封装。企业内部的装配、拧紧、码垛等具身操作 episode，不再只是“训练日志”或“实验数据”，而需要在统一 schema 与质量分级基础上，打包成具有清晰边界、可被发现和管理的“数据资产”，并通过标准化元数据映射到数据空间目录之中。

从粒度上看，操作数据资产可以有三个典型层级：一是以单个 episode 为单位，将某一次完整操作过程封装为最细粒度的资产，适合精细分析、策略对比和故障追溯；二是以生产批次或时间窗为单位，将一组 episode 聚合为“批次资产”，更贴近企业的生产与质量管理逻辑；三是以任务组或工艺段为单位，将沿工艺路线多个工序的 episode 串联为“任务组资产”，

便于在跨工序优化、端到端仿真或供应链协同中使用。无论采用哪一种粒度，对外暴露时都需要一个稳定的资产标识（asset_id），并能追溯到内部 episode_id、workorder_id、process_id 等业务主键，从而在不暴露内部系统细节的前提下，维持技术与业务之间的可追溯性。

围绕这些数据资产构建元数据模型时，一方面要满足国家对可信数据空间“可信可管、互联互通、价值共创”三类能力的要求，另一方面要与国际上以国际数据空间（International Data Spaces, IDS)/国际数据空间协会（International Data Spaces Association, IDSA）、Gaia-X、Catena-X 等数据空间实践中形成的目录与连接器模式相衔接。从字段维度看，一份可对接数据空间目录的操作数据资产元数据，至少应包含以下几类核心信息：其一是身份与来源信息，用于实现跨空间的资产发现与责任追踪；其二是技术特征与模态信息，为潜在使用方评估技术适配性提供依据；其三是质量与风险属性，为数据空间中的“高质量数据资源”识别和分级提供支撑；其四是使用条件与合规约束。通过这样一套“薄而全”的元数据建模，操作数据资产在进入数据空间之前就具备了可发现、可评估、可治理的基本条件，也为后续在连接器中绑定使用控制策略和审计规则奠定了结构基础。

54.2 “工业具身操作数据——可信数据空间”参考架构

在进一步给出参考架构之前，有必要将其与工业现场广泛存在的传统 MES/SCADA 数据链路作一横向比较。MES/SCADA 系统长期承担生产执行管理、设备状态监控、报警与处置、工艺过程追踪以及报表分析等职责，其优势在于对生产过程的可视化、可调度和可追溯管理，但其数据组织粒度通常围绕设备点位、工序事件、工单状态和统计结果展开，较少直接面向具身学习任务组织连续多模态时序数据。对于工业具身智能而言，仅依靠 MES/SCADA 链路通常难以形成以 episode 为核心、同时绑定轨迹、视觉、力觉、任务语义、工艺参数和质量结果的统一样本，也难以支持跨主体共享时所需的资产视图、使用控制和审计闭环。因此，本文提出的参考架构并不是替代 MES/SCADA，而是在保留其生产管理与过程追溯能力的基础上，补充面向多模态操作数据组织、资产化封装、可信流通与模型迭代的能力层。

表 4 传统 MES/SCADA 数据链路与面向可信数据空间的工业具身操作数据架构对比
Table 4 Comparison between traditional MES/SCADA data link and industrial embodied operation data architecture for trusted data space

比较维度	传统 MES/SCADA 数据链路	面向可信数据空间的工业具身操作数据架构
主要目标	生产执行、设备监控、报警处置、工单追踪、报表分析	多模态操作数据组织、策略学习支撑、跨主体可信共享与模型迭代
核心数据粒度	点位、事件、工单、工序、批次、报表	episode、批次资产、任务组资产
数据模态	结构化业务数据、设备状态、部分过程变量	本体轨迹、视觉、力觉/触觉、任务语义、工艺参数、质量结果等多模态组合
时间组织方式	以设备采样点、事件时间或业务记录时间为主	以统一时间基准组织连续时序，并强调跨模态对齐
与工艺/质量对接	通过工单、工序、检验记录管理	在 episode 内同时绑定工艺参数、步骤语义与质量结果
对训练的支持方式	更多支撑统计分析、追溯和报表	可形成主训练集、异常样本集、评估集和审计集

对外共享方式	多为系统对接或报表/接口导出	基于资产视图、连接器、目录、数字合约和使用控制进行共享
跨组织治理能力	通常局限于企业内系统集成	支持跨主体可控授权、审计留痕、用途约束和结果追踪
数据主权与收益分配	通常不作为核心设计目标	强调数据主权、使用边界、价值结算与责任链路

在资产化与元数据建模之上，可以构建一个面向“工业具身操作数据——可信数据空间”的参考架构，将国家行动计划提出的目标，与电气和电子工程师协会 (Institute of Electrical and Electronics Engineers, IEEE) Std 3158-2024 可信数据空间系统架构有机结合。在这一架构中，车间级具身智能单元被抽象为数据提供节点，工厂内部的操作数据资产库通过连接器与外部可信数据空间进行交互，连接器在数字合约与使用控制策略约束下完成数据请求的转译与响应，实现操作数据在符合国家制度与技术规范前提下的可控流通与可信共享。与传统 MES/SCADA 链路相比，这一架构的关键增量不在于替代生产执行与设备监控系统，而在于将原本分散在控制层、感知层和业务层的数据进一步组织为可训练、可评估、可共享的多模态 **episode** 资产，并通过连接器和合约机制将其纳入跨主体治理框架。

从工厂内部视角看，操作 **episode schema**、质量分级与数据目录共同构成了“共享准备层”。这一层通常部署在企业 IT 或数据中心侧，直接对接车间/工厂级处理平台产生的标准化 **episode** 样本和质量标签。一方面，它负责从内部资产库中选取符合共享前提的数据子集，例如仅选择质量等级为 A/B、已经完成脱敏处理、并通过内部合规审核的 **episode** 或批次；另一方面，它根据不同数据空间或使用场景的要求，对这些资产进行视图裁剪与聚合。这一层的输出，不再是原始日志，而是带有“对外视图定义”和“内外字段映射”的共享候选资产。

在可信数据空间系统架构中，‘连接器’是部署在数据提供方控制域内的客户端程序/用户节点组件：该组件在本地侧承接数据资产管理、数据资产控制与可信运行等功能，在外部侧通过标准化接口与第三方服务与提供方的平台互联（如数据资产管理平台、可信服务平台、数据综合服务平台等），从而实现面向可信数据空间的可控共享与可追溯使用。连接器主要承担三类职责：其一是数据资产与目录对接（供需对接），将企业内部完成生成、识别、维护的数据资产及其元数据纳入数据目录管理，并支撑条目共享与供需发布；同时配合合约管理完成共享对象的可交换描述与管理。其二是请求接入与策略协同，当外部数据使用方发起访问请求时，连接器负责完成客户端侧的身份识别与认证接入，并围绕合约协商与控制策略生成/确认等环节进行策略协同，支撑数据在传输与交付时携带并绑定相应控制策略。其三是使用控制执行与合规闭环，连接器在数据交付与使用全过程中执行控制策略、监测数据状态并按需触发数据删除，同时生成并管理日志、支撑审计，并为计量计费、清分结算提供可核验依据，形成‘策略—执行—记录—核验/结算’的闭环。需要强调的是，连接器承担的是数据交换、策略治理和合规控制职责，而不是毫秒级运动控制职责；对于对时延极度敏感的实时任务，外部数据空间链路应与现场控制闭环解耦，外部优化结果更适宜以模型包、参数包、规则包或推理容器的形式回流，经企业侧验证后再部署到本地边缘执行环境，而不宜通过远程在线调用直接介入机器人控制。

数字合约与使用控制策略是实现‘可信可管’的核心机制。使用控制通常基于开放数字权利语言 (Open Digital Rights Language, ODRL)等政策语言定义由‘允许、禁止、义务’组成的规则集合，并作为数据合约的一部分，由连接器进行技术层面的强制执行。结合可信数据空间发展行动计划中对‘可信管控能力’的要求，这些规则在工业具身操作数据场景中可以具体化为：仅允许将数据用于模型训练和评估，不得用于重新识别具体生产组织或操作员；

仅允许在约定的时间窗口内访问，超出期限自动失效；限制访问频次和数据量，防止大批量批出造成潜在泄露风险；要求使用方在本地存储时采用加密或访问审计，并在一定时间后执行删除或归档义务等。进一步地，对于工业具身操作数据，仅有静态访问控制仍然不足。数字合约还应对训练用途、衍生模型产物和结果输出边界作出约束，例如：是否允许跨项目复用训练得到的模型参数，是否允许提取或输出工艺敏感特征，是否允许将学习得到的隐式工艺知识迁移至其他客户场景，是否要求对训练过程、模型版本和输出结果进行审计留痕，以及当模型服务产生经济收益时，是否按约定规则对数据提供方进行收益确认或结算。对于可能暴露关键工艺参数、质量阈值、设备能力边界或配方信息的数据资产，还应结合字段裁剪、统计降精度、敏感区间模糊化、用途限定和结果审计等机制，降低通过模型学习反向推断企业核心 know-how 的风险。连接器在执行这些策略时，需要将策略与内部资产库的元数据结合起来，动态决定哪些字段可见、哪些 episode 可选、返回多大粒度的结果，以及在响应过程中记录哪些审计信息。

从外部可信数据空间的角度看，这一参考架构与 Catena-X 等工业数据空间中的“数据空间操作模型”是兼容的：企业通过自有连接器完成身份管理、合规检查和数据产品注册，数据空间则负责提供统一的目录、发现与匹配机制，以及跨参与方的合规与互信框架。对于工业具身操作数据而言，区别主要在于数据产品的结构与治理要求：一方面，数据产品需要基于 episode schema 和质量分级构建，强调对动作轨迹、多模态观测和质量结果的统一组织；另一方面，数字合约和使用控制必须显式考虑到 OT-IT 安全域划分、工艺敏感信息保护以及与国家相关政策和行业标准的一致性。

综上，本节给出的参考架构在逻辑上形成了自下而上的三层链路：车间与工厂内部通过标准化 episode 和质量控制构建操作数据资产库；共享准备层在企业域内完成资产视图裁剪与元数据建模；连接器在数字合约与使用控制策略的约束下，将这些资产安全地暴露给可信数据空间，实现跨主体、跨系统乃至跨行业的“可控共享”。

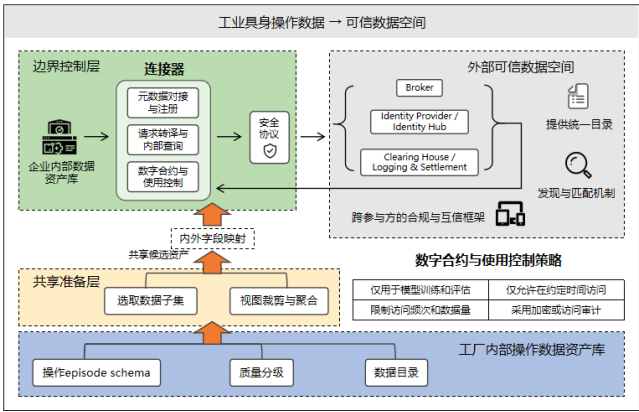


图 4 面向可信数据空间的工业具身操作数据共享与治理参考架构

Fig.4 Reference architecture for sharing and governance of industrial embodied operation data for trusted data space

54.3 工业案例与实现路径

本节以“某汽车零部件工厂的拧紧装配单元”为例，给出一个说明性工程案例，用于展示工业具身操作数据在采集、处理、共享、模型回流与边缘部署过程中的关键衔接关系。该工

厂在总成装配线上部署了一批具身智能拧紧单元，负责对关键螺栓进行多阶段拧紧作业，并希望通过与外部算法供应商合作，在可信数据空间中共享部分操作数据，以持续优化拧紧策略与异常检测模型，同时严格遵守企业内部的安全与合规要求。

在现场采集阶段，机器人控制柜以毫秒级周期向边缘网关输出关节角、末端位姿、控制指令和任务状态；智能拧紧控制器输出扭矩-角度曲线、目标扭矩、拧紧阶段信息及合格/不合格判定；相机通过工业以太网向边缘节点推送关键帧图像或编码视频；PLC 提供工位节拍、工单号、工序开始/结束信号等离散事件。边缘网关部署在 OT 区域内，只读接入机器人、拧紧控制器、相机和 PLC，基于统一配置完成多通道数据采集与时间戳标记，并在不影响实时控制的前提下，将原始信号缓冲为围绕单次拧紧动作的原始 episode 片段。

在内部处理阶段，车间级处理平台通过 PTP/NTP 等机制维护统一时间基准，将不同工位、不同设备上报的数据流对齐到同一时间轴上。平台根据工序开始/结束信号、拧紧控制器的“拧紧开始/完成”事件，以及机器人任务状态，将每次螺栓拧紧过程切分为一个独立的操作 episode。对于每个 episode，系统聚合机器人轨迹、扭矩-角度曲线、关键帧图像、力觉变化以及工单号、螺栓位点 ID、目标扭矩等业务标签，并利用标定管理服务，将相机与机器人之间的手眼标定、工装坐标系等信息映射到统一的空间坐标系中。同时，数据清洗与质量评估模块检测通讯丢包、传感器饱和、机器人急停、拧紧曲线异常等情况，为 episode 打上 A/B/C 三级质量标签：A 级表示传感与标注完备、无明显异常，可作为训练与评估主数据；B 级表示存在局部缺失或轻微异常，可用于鲁棒性训练；C 级则用于审计和问题定位，不进入算法训练主库。经过这一系列处理，工厂内部形成了围绕 episode schema 组织的“拧紧操作数据资产库”，并通过内部数据目录支持按工序、螺栓类型、质量结果等多维检索。

在资产化与对外视图设计阶段，工厂的数据治理团队会根据合作范围和合规要求，从资产库中筛选符合条件的 episode 子集。对于拧紧任务，这类对外视图通常包括标准化的机器人轨迹、扭矩-角度曲线、关键帧图像或局部视觉 patch、质量结果，以及经归一化处理的工艺参数，但不包含完整工单号、完整车辆识别号码 (Vehicle Identification Number, VIN) 号和细粒度生产排程等敏感信息。每一类视图对应一个或若干“数据产品”，在企业内部被赋予独立的 asset_id 和使用策略。

在通过数据空间共享的环节，工厂在其 IT/DMZ 区域部署符合可信数据空间规范连接器，将选定的数据产品及其元数据注册到所属行业数据空间中。元数据中不仅包含数据产品的技术特征，还包含数据主权与使用限制声明。数字合约以机器可读的使用控制策略形式附着在数据产品上，规定允许用途、禁止行为和必要义务。算法供应商通过自身连接器在数据空间目录中发现该数据产品后，发起访问请求，请求在空间内完成身份认证和合规检查，随后传输至工厂侧连接器。工厂连接器根据数字合约与内部策略，在本地将外部请求转译为对资产库的查询，动态裁剪字段与样本范围，并在响应过程中记录访问时间、调用方身份、数据量和用途声明等审计信息，最终以流式或批量方式向对方返回脱敏后的 episode 数据。

在外部算法回流环节，算法供应商利用获取到的拧紧 episode 数据，训练或微调具身策略模型和异常检测模型。模型训练完成后，供应商通过数据空间或其他约定渠道向工厂返回模型权重、推理容器或策略更新包，同时附带必要的说明文档和性能报告。需要强调的是，在此类“以数据换能力”的合作模式下，企业关注的不仅是访问是否合规，还包括模型是否学习并固化了超出约定范围的工艺知识。对于可能隐含关键拧紧阈值、材料特性、设备能力边界或质量判定规律的数据，不宜简单以原始 episode 全集对外开放，而应根据风险等级采用字段裁剪、敏感参数模糊化、分层抽样、用途限定和输出审计等方式控制暴露面；同时，应在合作合约中明确模型服务的使用范围、回流边界、跨项目复用限制、责任追踪方式以及收益确认机制，避免因第三方对大量 episode 的学习而反向推导企业核心工艺参数，造成事实上的技术失密和收益失衡。这里的“策略回流”应理解为版本化模型或参数的离线/近线

更新，而不是通过数据空间链路对现场控制器进行远程在线闭环干预。工厂内部的验证团队在隔离的测试环境中，用内部未共享的验证集对新策略进行再评估，确认其在安全边界内带来明显收益后，再逐步将新策略部署到车间边缘节点中，并在一段时间内以“灰度发布”的方式监控其表现。对于上线后的边缘执行环境，还应保留上一稳定版本、本地保底控制逻辑和安全互锁机制；一旦出现异常动作、时延抖动、外部链路不可用或性能退化，应能够快速回退到稳定版本或切换到本地保底策略。新的 episode 数据会继续按照前述流程进行采集、处理和资产化，其中一部分在未来再次进入数据空间，驱动新一轮的策略迭代，形成“现场采集—内部资产化—空间共享—外部优化—模型回流—再采集”的闭环。

在这一说明性工程案例中，可以看到前文提出的各个概念在工程上如何彼此衔接。该案例主要用于展示工业具身操作数据从现场采集、内部资产化、可信共享到模型回流和再部署的完整路径，以及其中涉及的安全边界、治理机制和系统组件关系。需要说明的是，本文在此并未将该案例作为对策略收益、质量改善或节拍优化效果的定量证明；相关定量验证仍需结合具体试点部署，围绕 episode 生成完整率、多模态对齐误差、共享链路时延、异常检测效果、策略回流后的质量改善与节拍变化等指标进一步开展。

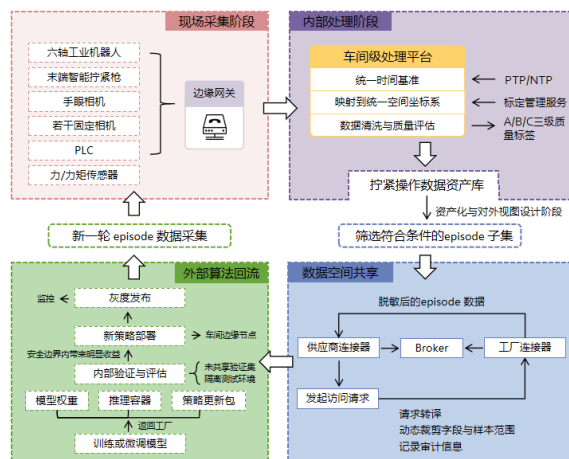


图 5 基于可信数据空间的拧紧装配操作数据闭环应用案例

Fig.5 Closed-loop application case of tightening assembly operational data based on trusted data space

65 结论与展望

本文围绕“工业具身智能操作数据如何从车间走向数据空间”这一核心问题进行了系统探讨。首先，提出了“四维分析框架”，并引入以操作 episode 为核心的要素模型与示例 schema。其次，对代表性具身操作数据集进行了对比分析，提炼出若干可直接迁移到工业场景的设计模式并识别关键缺口。再次，提出了“工业具身操作数据采集与处理参考架构”，以及面向“工业具身操作数据→可信数据空间”的共享与治理参考架构，并结合工程案例说明了其应用路径与工程衔接关系。未来工作将结合试点部署，进一步围绕数据质量、共享链路策略回流效果建立量化评估指标体系，并开展定量验证。未来工作将从跨企业、多数据空间的协同机制、策略的在线更新与数据闭环、安全与隐私评估及标准化工作对接等方面进一步扩展；同时，将进一步探索面向工业 SOP、工艺卡和作业指导书的语义发现方法，利用大语言模型将非结构化作业文本转化为任务图、阶段标签乃至线性时序逻辑 (Linear Temporal Logic,

LTL)规约，并与 episode 标注和质量约束联动，以提升长序列操作数据的逻辑一致性表达能力和高层语义建模能力。

本文把散落于具身智能、工业互联网和数据空间三个领域的零散做法，重组为一套连贯的问题表述、比较坐标系和工程参考架构。希望这些工作能够为后续的理论研究、工程实践和标准制定提供一个相对稳定的出发点，推动工业具身智能操作数据在“可用、好用”的同时，也在“可信、可管”的轨道上发展。

参考文献：

- [1] PFEIFER R, IIDA F. Embodied artificial intelligence: Trends and challenges[M]//Embodied Artificial Intelligence. Berlin, Heidelberg: Springer, 2004: 1-26.
- [2] PFEIFER R, BONGARD J. How the Body Shapes the Way We Think: A New View of Intelligence[M]. Cambridge, MA: MIT Press, 2006.
- [3] 金立民, 王海超, 谷江春, 等. 低空具身智能频谱管控技术研究[J]. 数据采集与处理, 2025, 40(1): 45-55.
JIN Limin, WANG Haichao, GU Jiangchun, et al. Research on Low-Altitude embodied artificial Intelligence-Enabled spectrum management and control technology[J]. Journal of Data Acquisition and Processing, 2025, 40(1): 45-55.
- [4] 孙一铭, 赵柯嘉, 王硕, 等. 面向无人机的低空视觉数据集研究综述[J]. 数据采集与处理, 2025, 40(2): 274-302.
SUN Yiming, ZHAO Kejia, WANG Shuo, et al. Research review on low-altitude visual datasets for unmanned aerial vehicles[J]. Journal of Data Acquisition and Processing, 2025, 40(2): 274-302.
- [5] DUAN J, YU S, TAN H L, et al. A survey of embodied AI: From simulators to research tasks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(2): 230-244.
- [6] DU Y. A Comprehensive Survey on Embodied AI[EB/OL]. arXiv:2407.06886, 2024.
- [7] 高超, 杨莹, 陈世超, 等. 多模态模型驱动的具身智能研究综述[J]. 智能感知工程, 2025, 2(2): 1-12.
GAO Chao, YANG Ying, CHEN Shichao, et al. Survey on multimodal model-driven embodied intelligence research[J]. Intelligent Perception Engineering, 2025, 2(2): 1-12.
- [8] 王文晟, 谭宁, 黄凯, 等. 基于大模型的具身智能系统综述[J]. 自动化学报, 2025, 51(1): 1-19.
WANG Wensheng, TAN Ning, HUANG Kai, et al. Embodied intelligence systems based on large models: A survey[J]. Acta Automatica Sinica, 2025, 51(1): 1-19.
- [9] 徐凯, 赵航, 胡瑞珍, 等. 面向柔性制造的具身智能综述[J]. 机器人, 2025, 47(4): 581-624.
XU Kai, ZHAO Hang, HU Ruizhen, et al. Embodied intelligence for flexible manufacturing: A survey[J]. Robot, 2025, 47(4): 581-624.
- [10] 中国信息通信研究院. 智能机器人技术产业发展白皮书（2023）[R]. 北京: 工业互联网产业联盟(AII), 2023.
- [11] 袁栩聪, 张俊雅. 2025 年具身智能行业研究: 跨领域融合引领的新一轮智能革命[R]. 2025.
- [12] 国家数据局. 可信数据空间发展行动计划（2024—2028 年）[EB/OL]. 国数资源（2024）119 号, 2024.
- [13] XU Z, WU K, WEN J, et al. A survey on robotics with foundation models: Toward embodied AI[EB/OL]. 2024: arXiv: 2402.02385. <https://arxiv.org/abs/2402.02385>.
- [14] FIROOZI R, TUCKER J, TIAN S, et al. Foundation models in robotics: : Applications, challenges, and the future[J]. International Journal of Robotics Research, 2025, 44(5): 701-739.
- [15] COLLABORATION E, O'NEILL A, REHMAN A, et al. Open X-embodiment: Robotic learning datasets and RT-X models[EB/OL]. 2023: arXiv: 2310.08864. <https://arxiv.org/abs/2310.08864>
- [16] KAWAHARAZUKA K, MATSUSHIMA T, GAMBARDELLA A, et al. Real-world robot applications of foundation models: A review[J]. Advanced Robotics, 2024, 38(18): 1232-1254.
- [17] KHAZATSKY A, PERTSCH K, NAIR S, et al. DROID: A large-scale in-the-wild robot manipulation dataset[EB/OL]. 2024: arXiv: 2403.12945. <https://arxiv.org/abs/2403.12945>

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

[19] BAI S, SONG W, CHEN J, et al. Towards a unified understanding of robot manipulation: A comprehensive survey[EB/OL]. 2025: arXiv: 2510.10903. <https://arxiv.org/abs/2510.10903>

[20] FANG H S, FANG H, TANG Z, et al. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan. IEEE, 2024: 653-660.

[21] 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.

[22] YAO X, ZHOU H, MEES O, et al. Bridging language and action: A survey of language-conditioned robot manipulation[EB/OL]. 2023: arXiv: 2312.10807. <https://arxiv.org/abs/2312.10807>

[23] NINOV N N, ASIF W. Language Conditioned Robotics: Current Trends and Future Directions—A Survey Paper[R]. SSRN 5073634, 2024.

[24] YOUNAN M, HOUSSEIN E H, ELHOSENY M, et al. Challenges and recommended technologies for the industrial Internet of Things: A comprehensive review[J]. Measurement, 2020, 151: 107198.

[25] HORAK T, STRELEC P, KEBISEK M, et al. Data integration from heterogeneous control levels for the purposes of analysis within industry 4.0 concept[J]. Sensors, 2022, 22(24): 9860.

[26] ESCOBAR C A, MCGOVERN M E, MORALES-MENENDEZ R. Quality 4.0: A review of big data challenges in manufacturing[J]. Journal of Intelligent Manufacturing, 2021, 32(8): 2319-2334.

[27] RAJNOHA R, HADAC J. Strategic key elements in big data analytics as driving forces of IoT manufacturing value creation: A challenge for research framework[J]. IEEE Transactions on Engineering Management, 2024, 71: 90-105.

[28] ALI MIRANI A, VELASCO-HERNANDEZ G, AWASTHI A, et al. Key challenges and emerging technologies in industrial IoT architectures: A review[J]. Sensors, 2022, 22(15): 5836.

[29] INOUE T, MORITA T. Trusted data space technology for data governance in the IOWN era[J]. NTT Technical Review, 2023, 21(4): 55-59.

[30] LIANG Z Y, LIU G Y, REN Y, et al. Trustworthy data space collaborative trust mechanism driven by blockchain: Technology integration, cross-border governance, and standardization path[J]. Information, 2025, 16(12): 1066.

[31] ZVEI, VDMA, Fraunhofer ISST. Manufacturing-X Data Space Study[R/OL]. Frankfurt am Main, 2023.

[32] MANOURY M, RIEDELSHEIMER T, HELLMEIER M, et al. Supporting changes in digital ownership and data sovereignty across the automotive value chain with catena-X[J]. Procedia Computer Science, 2025, 253: 374-383.



作者简介:

韦莎 (1985-), 通信作者, 女, 博士, 副总工程师, 正高级工程师, 研究方向: 智能制造、数字孪生和可信数据流通研究, E-mail: weisha@caict.ac.cn。



费恩琪 (2004-), 女, 本科生, 研究方向: 图像篡改定位。

度强化学习与嵌入式、具身智能、智能体、机器人、数字孪生等。

孙闻 (1989-), 男, 博士后, 研究方向: 数字孪生, 工业智能体, 智能机器人等。

汤丰恺 (2001-), 男, 工程硕士, 研究方向: 人工智能、深

殷泽明 (1996-), 男, 硕士, 研究方向: 机器人与人工智能、



智能设备数据采集、具身智能控制、智能体行为分析与可靠性验证等相关技术。

Acquisition, Processing and Trusted Sharing of Operational Data for Industrial Embodied Intelligence

WEI Sha¹, FEI Siqi², SUN Chuang¹, TANG Fengkai¹, DUAN Zeming¹

(1. Institute of Informatization and Industrialization Integration, China Academy of Information and Communications Technology, Beijing 100191, China; 2. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: With industrial embodied intelligent equipment, such as intelligent robots, being rapidly deployed on the shop floor, high-quality multimodal operational data has become a critical bottleneck for policy learning, robust planning, and online optimization. Existing public datasets are largely collected in laboratory or simulated environments, leaving systematic gaps to industrial settings in terms of process constraints, hardware heterogeneity, operational safety, and data security/compliance. To address this, this paper proposes a structured modeling framework for industrial embodied-intelligence operational data that spans data types–acquisition architecture–processing pipeline–sharing mechanism, and summarizes key engineering steps including multi-source data organization and key processing steps. Furthermore, aligned with the Trusted Data Space Development Action Plan (2024–2028) and its three core capabilities of trusted governance and control of data, resource interaction, and value co-creation, we develop a reference architecture based on the trusted data spaces system. By abstracting workshop-level embodied-intelligence units as data nodes and leveraging edge-side pre-processing, connectors, data catalogs, and digital contracts with usage control, the proposed solution enables controllable authorization, real-time evidence preservation, and traceable usage across organizations, thereby supporting trusted sharing and reuse of operational data. Based on literature review and engineering cases, we outline a practical implementation path, providing references for industrial data-solution design and standardization toward trusted data spaces.

Highlights:

1. Proposes a four-dimensional analytical framework (data types, acquisition architecture, processing pipeline, and sharing mechanism) specifically for industrial embodied intelligence operational data, addressing the gaps in existing reviews regarding industrial data characterization.
2. Introduces an operation episode-centric element model and a standardized schema that integrates multi-modal trajectories, task semantics, process parameters, and quality labels, bridging the gap between algorithm training and industrial production management.

3. Constructs a hierarchical reference architecture for data acquisition and processing, translating scattered OT signals from robotic controllers, sensors, and PLCs into structured, quality-graded operational samples.
4. Designs a trusted sharing and governance architecture aligned with the national "Trusted Data Space" initiative, enabling cross-organizational controlled authorization, usage control, audit trails, and result traceability for operational data.

Keywords: industrial embodied intelligence, operational data, data acquisition and processing, multimodal datasets, and trusted data spaces