

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS 018—2024

具身智能语料库建设导则

Construction Guidelines for Corpus of Embodied Intelligence

2024 - 11 - 16 发布

2024 - 11 - 18 实施

目 次

前 言 II

引 言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 2

5 具身智能数据 2

 5.1 概述 2

 5.2 一般原则 2

 5.3 数据集 2

 5.4 数据资源 3

6 具身智能数据采集 7

 6.1 概述 7

 6.2 真实数据采集 7

 6.3 合成数据采集 8

 6.4 数据资源提交方式 9

7 具身智能语料生产 9

 7.1 采集数据完整性检查 9

 7.2 数据资源清洗过程 9

 7.3 数据资源标注过程 10

8 数据安全 11

 8.1 安全原则 11

 8.2 安全性评价 11

 8.3 制度 11

 8.4 技术 12

 8.5 人员 12

附 录 A （资料性） 具身智能大模型方案 13

 A.1 语言大模型 13

 A.2 视觉语言大模型 13

 A.3 视觉语言动作大模型 14

参 考 文 献 16

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：上海库帕思科技有限公司、人形机器人（上海）有限公司、复旦大学、上海交通大学、同济大学、上海人工智能创新中心、上海智元新创技术有限公司、上海傅利叶智能科技有限公司、上海开普勒机器人有限公司、中国电子科技集团公司第二十一研究所、星海图（苏州）人工智能科技有限公司、上海钛米机器人股份有限公司、松应科技有限公司、四川天链机器人股份有限公司、上海市先导产业促进中心、工业互联网创新中心（上海）有限公司、上海中创产业创新研究院、国创智造科技（上海）有限公司。

本文件主要起草人：山栋明、黄海清、许彬、姜育刚、卢策吾、何斌、庞江淼、闫维新、高宇翔、陈春玉、王洪武、郑忠斌、高平、施佳樑、卢恒、张裕珍、江磊、刘宇飞、黄媚、刘传厚、陈曦、沈荣港、冯瑞、蒋烁、王志鹏、王宇、姚卯青、朱建超、胡硕、蔡宇圣、沈林鹏、陈宇涛、李昀佶、许华哲、虞磊、潘晶、沈满、聂凯旋、张小波、黄巍、颜进、郑理莹、宋琮、李琰、方金武、王道洋、蒋龙泉、朱立奇、贺仁龙、郑茂宽。

首期执行单位：上海库帕思科技有限公司、人形机器人（上海）有限公司、上海人工智能创新中心、上海智元新创技术有限公司、上海傅利叶智能科技有限公司、上海开普勒机器人有限公司、中国电子科技集团公司第二十一研究所、星海图（苏州）人工智能科技有限公司、上海钛米机器人股份有限公司、松应科技有限公司、四川天链机器人股份有限公司。

本文件版权归上海市人工智能行业协会所有。未经许可，不得擅自复制、转载、抄袭、改编、汇编、翻译或将本标准用于其他任何商业目的。

引 言

人工智能是新一轮科技革命和产业变革的重要驱动力量，语料数据则是人工智能研究和应用不可或缺的资源，高质量的语料库更是人工智能赋能新质生产力的关键。

在人工智能的浪潮中，具身智能作为前沿科技的代表，正受到国家和上海市的高度重视。《具身智能语料库建设导则》的编纂，正是在这一背景下应运而生，旨在为该领域的发展提供坚实的数据基础和标准化指导。

国家层面，工信部发布的《人形机器人创新发展指导意见》强调了开展人形机器人标准化路线图研究的重要性，并提出建立健全人形机器人产业标准体系。这不仅为具身智能语料库的建设提供了政策支持，也为行业的健康发展指明了方向。

上海市作为科技创新的先行者，率先发布全国首批人形机器人具身智能标准，包括《具身智能智能化等级分级指南》和《人形机器人分类分级应用指南》。这些标准不仅规范了人形机器人的技术语言和发展路径，也为语料库建设提供了明确的技术等级划分依据。上海还计划建设完成多地联动的具身智能训练场，支持超100台异构机器人采集训练，这将为具身智能的语料库建设提供丰富的实践场景和数据资源。

此外，上海市人民政府印发的《上海市进一步推进新型基础设施建设行动方案（2023-2026年）》中提到，将布局智能机器人创新基础设施，建设“大模型+人形机器人”协同创新平台，这为具身智能语料库的建设提供了政策和技术支持。

本文件的编写，紧密结合这些最新的政策导向，为语料库的建设提供科学、系统、标准化的指导。本导则将详细阐述如何收集、标注和利用数据，以满足具身智能系统在不同场景下的需求，同时确保数据的质量和多样性。通过遵循本导则，我们期待能够推动具身智能技术的创新和应用，为具身智能产业的科学健康发展注入强大动力。

具身智能语料库建设导则

1 范围

本文件提供了建设具身智能模型训练数据内容、数据采集方式、语料生产方法和数据安全方面的技术指导方法。

本文件适用于具身智能语料库的研究、开发、维护、应用、评估等工作。其它与具身智能语料库建设相关的工作也可参照使用。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

- GB/T 36073-2018 数据管理能力成熟度评估模型
- GB/T 35273-2020 信息安全技术 个人信息安全规范
- GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求
- GB/T 4894-2009 信息与文献 术语
- GY/T 360-2022 广播电视和网络视听节目内容标识标签规范
- GY/T 353-2021 网络视听节目视频格式命名及参数规范
- YD/T 4245-2023 电信网和互联网数据脱敏技术要求和测试方法
- T/SAIAS 015—2024 语料库建设导则

3 术语和定义

T/SAIAS 015—2024界定的以及下列术语和定义适用于本文件。

3.1

数据资源 data resources

以电子化形式记录和保存的具备原始性、可机器读取、可供社会化再利用的数据集合。

[来源：GB/T 4894-2009，4.7.3.2.3，有修改]

3.2

本体 Embodied intelligent ontology

一般指在物理环境或者虚拟环境中进行感知和任务执行的机器人本体。

3.3

数据片段 data clip

以片段形式记录和保存的具身智能模型训练使用的数据集合。

注：是按照一定目的和方法进行选择并有序排列的数据汇集。

3.4

数据时刻 tick dataset

在同一时刻中，由各相机视频帧、雷达数据、本体数据、轨迹导航数据组成的数据集合。

3.5

模拟仿真平台 simulation platform

基于计算机的系统，用于模拟现实世界或想象中的环境、条件或系统。

3.6

具身智能 Embodied AI

基于物理身体进行感知和行动的智能系统，其通过智能体与环境的交互获取信息、理解问题、做出决策并实现行动，从而产生智能行为和适应性。

注：本标准限定的具身智能本体包括泛人形机器人、仿人机器人、智能机器人等。

4 缩略语

下列缩略语适用于本文件。
NSFW 不适于工作场所浏览 (Not Safe For Work, Not Suitable For Work)
FOV 景深 (Field Of View)
SFT 有监督微调 (Supervised Fine-Tuning)

5 具身智能数据

5.1 概述

本章规定了具身智能语料库与作为语料库素材的数据资源的规范性说明。这些说明将帮助这些数据的使用方（本文件中的用户）了解数据的特性、质量和可用性，以对数据的应用价值作出初步评估。
本文件说明的语料库一般用作具身智能大模型训练使用，附录A说明了具身智能大模型方案。

5.2 一般原则

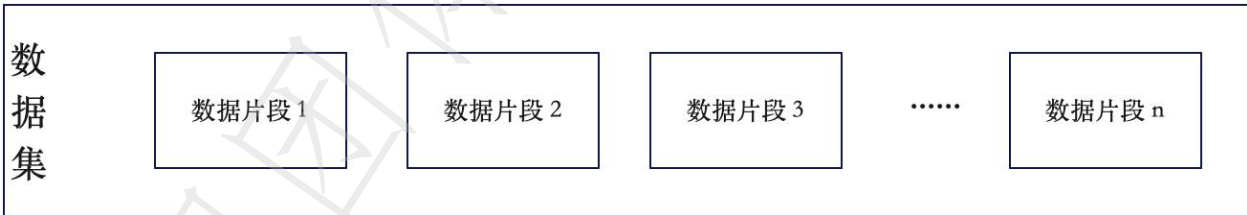
具身智能语料库与作为语料库素材的数据资源考虑的主要原则包括：

- a) 数据模态：具备多模态数据，并能涵盖多模态传感数据；
- b) 数据同步性：具备同步性，多模态多传感器数据的开始时间与结束时间是一致的，保持时空同步性；
- c) 数据规模：具有足够的规模，并能覆盖不同领域和不同应用场景；
- d) 数据多样性：来自多场景全时空数据；
- e) 数据密级：最小单位的文件不设置对其内容操作的权限控制；
- f) 数据质量：具有具有规范性、一致性、完整性、及时性、准确性、有效性和时效性，并不含水印和不适宜内容。

5.3 数据集

具身智能语料库数据集，由一系列数据片段的形式存储，由一系列数据片段有序或自由组合而成，适用于具身智能模型训练使用。数据集与数据片段关系如图1所示。

图 1 数据集与数据片段关系



5.3.1 数据片段

数据片段由连续的若干个数据时刻组成。数据片段与数据时刻的关系如图2所示。

图 2 数据片段与数据时刻的关系



5.3.2 数据时刻

数据时刻内容包含同一时刻的视频数据、雷达数据、轨迹导航数据以及本体数据，每一个数据时刻是时空同步的。

示例：按每秒 10 帧为例，一个数据时刻是在 1 帧内所有传感器数据的集合，即所有传感器在 100 毫秒（ms）内产生数据的集合。

数据时刻内容包括以下数据类型：

- a) 相机数据：指的是各相机数据；
- b) 雷达数据：指的是激光雷达数据、毫米波雷达数据、超声波等大数据等；
- c) 导航数据：指的是GPS、地图数据、GNSS数据、IMU数据等；
- d) 本体数据：指的是本体状态数据、控制命令数据、本体动态数据等；
- e) 标注数据：文本标注数据、图像标注数据、音频标注数据等；
- f) 合成数据：指的是通过仿真平台合成的数据；
- g) 其他数据：指的是传感器标定数据、文本数据、音频数据等。

5.4 数据资源

5.4.1 数据资源类型

5.4.1.1 视频数据

视频数据以一系列序列帧的形式存储，用于视频内容的存储和处理。具身智能视频类型如表1所示。

表 1 视频数据的指标

序号	指标	说明
1	视频类型	见表2
2	数据资源内容	视频及对应文字说明或视频介绍，包括拍摄地点、时间、设备、照明条件、天气条件、环境条件等信息
3	视频分辨率	宜1080P（1920x1080像素）及以上
4	视频帧率	宜25帧/秒及以上
5	单一视频时间长度	10秒及以上
6	视频文件格式	MP4/AVI/YUV/H264等
7	视频验收标准	除清晰外，对视频素材的随机抽样中，宜有80%包含主体（不含主体的视频素材示例，包括但不限于航拍、延时风景摄影等）

视频数据来自本体头部、胸口、左右手臂末端、第三方视角摄像头的视频序列，用于物体检测、语义分割、环境检测、场景设施识别等。摄像头视频数据分类及说明如表2所示。

表 2 视频数据分类表

序号	类型	说明
1	头部视频	主要指头部摄像头采集的视频数据
2	胸口视频	主要指胸口摄像头采集的视频数据
3	左右手臂末端视频	主要指左右手臂末端摄像头采集的操作目标近景视频数据
4	第三方视角视频	主要指从第三方视角采集的视频数据
5	红外视频	主要指使用红外摄像头拍摄，适用于夜间或光线不足的环境
6	立体视频	主要指采用两个或以上的摄像头从不同角度拍摄，用于深度感知和3D建模

视频数据还包含的天气条件、照明条件和场景类型等数据语义特征，具身智能数据语义特征如表3所示。

表 3 视频数据语义特征表

序号	类型	说明
1	天气	主要包括晴朗、多云、阴天、雨天、雾天、雪天、沙尘暴、雷电
2	时间	主要包括白天、夜晚、黎明、黄昏
3	光照	主要包括光线较好、光线较弱、黑暗、过曝/眩光、大面积阴影、反射虚影
4	场景	主要包括工厂、家居、办公室、医院、学校、商超、开放道路等。

相机视角与具身智能数据集保持一致，多个视角内的目标可追踪，保持特征一致。

5.4.1.2 雷达数据

雷达数据包括激光点云数据和毫米波数据。

激光点云数据可发射激光束并接收反射回来的信号，精确测量物理的距离、形状和位置，构建三维环境模型。激光点云数据的指标如表4所示。

表 4 激光点云数据的指标

序号	指标	说明
1	分辨率	宜达到车规级激光雷达的行业标准
2	精度	在±2cm以内
3	点云密度	宜每平方米90个点及以上，均匀分布
4	检测范围	宜30米及以上
5	帧率	宜10Hz及以上
6	动态范围	宜10厘米-100米
7	数据资源说明	点云数据及对应文字说明或介绍
8	噪声水平	距离误差±3cm以内，角度误差±1度以内，时间同步误差±10微秒以内
9	单一点云数据大小	宜50MB及以上
10	文件格式	PCD/LAS/XYZ
11	验收标准	除清晰外，对素材的随机抽样中，点数足够，分布均匀，且精度误差小，噪声水平低

相关激光点云数据需满足以上指标，若出现以下情形，则不适合作为激光点云数据，不适合作为激光点云数据如表5所示。

表 5 不适合作为点云数据资源的情形

序号	情形	说明
1	数据不完整	主要指缺少关键区域的点，包括目标物体、环境设备
2	分辨率过低	主要是指点云过于稀疏，无法捕捉物体细节
3	精度不高	主要是指点的位置信息误差大
4	噪声过多或同步性差	主要指包含大量错误、异常值或者多传感器数据无法准确同步
5	动态范围小	主要指无法覆盖远近距离的物体
6	反射强度不一致	主要指反射强度信息不稳定或不可靠
7	光照影响大	主要指强光或逆光条件下产生阴影或反射
8	天气影响严重	主要指雨、雪、雾等恶劣天气条件下的数据
9	遮挡严重	主要指其他物体遮挡无法获取完整目标物体视角
10	数据过时	主要指与当前环境变化较大，无法反映当前场景
11	法律和隐私问题	主要是指包含敏感信息，包括车牌号、人脸等
12	冗余数据	主要指大量重复或非常相似的数据

毫米波数据，擅长测量目标的距离和速度，对金属物理比较敏感，在恶劣环境下也能较好工作。毫米波数据资源的指标如表6所示。

表 6 毫米波数据的指标

序号	指标	说明
1	频率范围	宜76GHz及以上
2	检测范围	宜30米及以上
3	角分辨率	$\pm 0.5^{\circ}$ （长距离）/ $\pm 1^{\circ}$ （中距离）
4	距离分辨率	宜0.1米及以上
5	速度分辨率	宜0.1米/秒
6	更新率	宜20 Hz及以上
7	调制方式	FMCW
8	检测精度	距离： $\pm 0.05\text{m}$ ，角度： $\pm 0.5^{\circ}$ ，速度： $\pm 0.1\text{ m/s}$
9	文件格式	支持CAN/Ethernet等通信协议
10	验收标准	除清晰外，对素材的随机抽样中，分辨率达标，宜有80%包含主体。

5.4.1.3 音频数据

声音表征的数据，简称音频数据，是以声音波形的形式存储，通常以采样率和位深来表征声音信号。音频数据资源（或可再加工音频数据产品）的指标如表7所示。

表 7 音频数据的指标

序号	指标	说明
1	语种	汉语（普通话、粤语、沪语等）、英语（英国、美国）、阿拉伯语、俄语、日语
2	主题领域	参照《广播电视和网络视听节目内容标识标签规范》（GY/T 360-2022）中的“内容类内容特征子类别规范词”
3	数据资源内容	音频及对应文字说明或音频介绍
4	音频采样率	不小于44.1KHz
5	通道数	双声道/单声道（由原始资料特性决定）
6	单一音频时间长度	60秒及以上
7	量化精度	不低于16位
8	音频文件格式	WAV
9	音频验收标准	对音频素材的随机抽样中，宜有95%不存在表8中所示情况

相关的音频技术指标宜符合GY/T 353-2021的规定。

音频数据资源中包含了资源特征性描述的，相关的标识和（或）标签则宜满足GY/T 360-2022的规定。

音频数据需满足以上指标，若出现以下情形，则不适合作为音频数据，如表8所示。

表 8 不适合作为音频数据资源的情形

序号	情形	说明
1	文件噪音不合格	整段音频伴有严重的噪音
2	文件内容不合格	音频文件不包含目标类别的音频或含有误导性信息的音频
3	其它不合格	回音很大、过载严重

5.4.1.4 本体数据

本体数据包括本体状态数据、控制命令、诊断数据、本体动态数据以及传感器标定数据。本体数据资源的指标如表9所示。

表 9 本体资源指标

序号	指标	说明
1	类别	详见表14
2	数据资源内容	本体数据及对应文字说明
3	格式	JSON或数据库

4	验收标准	本体数据需准确，包含时间数据，且与视频数据保持时空一致性
---	------	------------------------------

本体信息数据资源分类及说明如表10所示。

表 10 本体信息数据资源分类表

序号	类型	说明
1	本体状态数据	主要包括启动/关闭、传感器健康状态
2	控制命令	主要包括转向、加速、制动、运动
3	诊断信息	主要包括本体故障代码和诊断信息
4	本体动态数据	主要包括速度、加速度、转向角度、制动力度、横向加速度、纵向加速度、轮速、本体位置、倾角、本体各关节电机功率

传感器标定数据在具身智能训练中能够误差校正，坐标系统统一以及确保不同传感器之间的数据关联性使传感器融合，提高具身智能整体性能。传感器标定数据资源指标如表11所示。

表 11 传感器标定数据指标和标准

序号	指标	说明
1	常见信息	相机内参、相机外参、激光外参、毫米波外参、INS的内参、INS的外参等
2	参数规定	参数宜与对应本体一一绑定，且定期重新标定进行维护
3	数据资源内容	传感器标定数据及对应文字说明
4	格式	JSON或采集数据包中
5	验收标准	根据各传感器参数，通过采集数据进行投影，根据重叠效果判断标定准确度

5.4.1.5 轨迹导航数据

轨迹导航数据包括定位数据、轨迹数据等。轨迹导航数据的指标如表12所示。

表 12 轨迹导航数据资源指标

序号	指标	说明
1	类别	详见表13
2	数据资源内容	轨迹导航数据及对应文字说明
3	格式	JSON或采集数据包
4	验收标准	定位精度符合传感器精度标准，包含时间数据，且与视频数据保持时空一致性

轨迹导航数据资源分类如表13所示。

表 13 轨迹导航数据资源分类

序号	类型		说明
1	定位数据	GPS	主要包括定位精度、定位连续性、时间戳同步
2		IMU	主要包括角速度零偏稳定性、加速度零偏稳定性、时间戳同步
3	轨迹数据	轨迹点	主要包括位置精度、速度精度、时间戳同步
4		轨迹预测	主要包括预测时间范围、预测频率、预测精度（包括minADE、minFDE）

5.4.2 数据信息存储格式

5.4.2.1 传感器标定信息格式

数据采集传感器标定的信息应包含：

- a) 传感器设备序列号及名称；
- b) 传感器的类型及数量；
- c) 传感器采集频率；
- d) 传感器内外参数。

5.4.2.2 元数据格式

采集的元数据宜包含以下信息：

- a) 通用数据：

- 1) 时间戳;
- 2) 唯一 ID。
- b) 本体数据:
 - 1) 类型;
 - 2) 关节位姿数据。
- c) 相机传感器数据:
 - 1) 彩色图像数据;
 - 2) 深度图像数据;
 - 3) 采样频率;
 - 4) 图像尺寸。
- d) 雷达传感器数据:
 - 1) 采集范围;
 - 2) 采集精度;
 - 3) 采样频率。
- e) 触觉传感器数据:
 - 1) 灵敏度;
 - 2) 分辨率;
 - 3) 响应时间。
- f) 音频传感器数据;
- g) 力觉传感器数据;
- h) 嗅觉传感器数据;
- i) 采集环境信息 (包括场景布局、地点、设备/道具布置情况等);
- j) 采集任务描述信息 (包括文字指令和对话信息等)。

5.4.2.3 元数据组帧格式

元数据组帧对齐后包含:

- a) 帧索引序列;
- b) 传感器数据序列;
- c) 本体数据序列。

6 具身智能数据采集

6.1 概述

本章规定了对语料库素材的数据资源的采集的规范性说明。采集内容包括真实数据采集和合成数据采集。这些说明对这些数据的采集方的采集介质、采集行为提供指导性意见,对数据进行统一标准。

6.2 真实数据采集

6.2.1 采集本体

本节规定了采集本体的基本配置,帮助采集方有效改装采集本体,保障采集数据一致性。

采集本体的相机数量为常见数量,即6个FOV相机,具体可根据实际采集需要进行加装。采集样本硬件如表14所示。

表 14 采集本体硬件

序号	硬件组件	关键参数	安装位置
1	头部相机	1颗 FOV30@8MP+, 1颗 FOV100+@8MP+	头部
2	胸口相机	1颗 FOV100@2.3MP+	胸口
3	左右手臂末端相机	2颗 FOV100@2.3MP+	左右手
4	第三方视角相机	1颗 FOV195@1.2MP+	第三方
5	存储系统	支持全量满带宽数据存储,每台配备4T固态硬盘	与操作系统算力、带宽不共用

采集本体软件如表15所示。

表 15 采集本体软件

序号	软件标准	描述
1	多模态传感器同步	支持多模态传感器时间同步与空间同步
2	全量满带宽数据存储	支持全量满带宽数据存储
3	OTA软件策略更新	支持通过OTA进行软件策略更新

6.3 合成数据采集

6.3.1 概述

虚拟合成数据通过基于世界模型的模拟仿真平台等工具生成的数据。

6.3.2 合成数据

为了确保仿真合成数据的有效性和适用性，考虑的主要因素包括：

- a) 高保真度：仿真环境尽量模拟现实世界中的物理和视觉特性，保证数据的真实性和准确性；
- b) 多样性：合成数据包含不同场景、多种任务和多种传感器数据，以确保模型的泛化能力；
- c) 可重复性：仿真数据生成可重复，允许在相同条件下生成一致的数据集；
- d) 同步性：多传感器采集的数据严格保持时间同步，特别是在多模态数据（包括图像、深度、雷达、IMU）融合时，时间戳精度足够高；
- e) 标准化格式：所有数据符合预定的标准格式，以便于存储、共享和分析；
- f) 随机化控制：为了增强数据的鲁棒性，合成数据进行随机化处理，包括视觉随机化（包括光照、材质等）和物理随机化（包括摩擦、重力等），以模拟不同环境下的变化。

6.3.3 合成数据生成方式

仿真合成数据生成包括以下步骤：

- a) 场景构建：使用3D数字孪生技术对真实场景进行虚拟化，构建高度逼真的仿真场景。场景的视觉和物理特性根据应用场景需求进行细致建模；
- b) 任务设计：机器人在场景中执行多种任务（包括抓取、移动、操作等），通过仿真平台模拟这些任务，并生成对应的动作和感知数据；
- c) 多传感器采集：通过机器人内部和外部的多种传感器（包括RGB相机、深度传感器、激光雷达等）同步采集数据。确保传感器数据的时间同步性，尤其在多模态融合时，数据采集频率需保持一致；
- d) 数据随机化：通过视觉和物理属性的随机化，生成多样化的数据集，以增加模型的鲁棒性和泛化能力。视觉随机化涉及材质、光照、视角等，物理随机化处理摩擦、重力等；
- e) 数据处理与存储：采集的数据宜经过适当的预处理，包括去噪、时间同步等步骤，随后以规定格式存储。文件名和目录结构须明确标识场景、时间和数据类型。

6.3.4 合成数据用途

仿真合成数据在具身智能系统中的应用场景广泛，主要包括以下几个方面：

- a) 数据泛化，仿真数据通过物品泛化和场景泛化来模拟各种现实环境中的变化情况。通过引入不同的物体类型、材质、大小、位置以及各种场景属性（包括光照、天气、地形等），仿真数据能够为模型提供更丰富的训练样本，确保模型在实际应用中具备更强的适应性和泛化能力。包括模型可以在虚拟环境中学会处理不同形状和尺寸的物体，从而在真实场景中也能表现良好；
- b) 极端情况（Corner Case）生成，解决现实世界中极端情况（包括罕见的物体交互、复杂的光照或极端天气等）的采集难题。仿真平台能够生成大量极端情况数据，这些情况在真实环境中难以捕捉，但却是模型泛化和鲁棒性的重要测试场景。通过系统地生成这些难以采集的极端案例，模型能够更好地处理特殊或意外的情境，从而在多变和复杂的环境中表现得更加可靠；
- c) 跨传感器/本体数据迁移，仿真合成数据的另一个重要用途是跨传感器和跨本体的数据迁移。通过充分利用已有数据，模型可以在不同的传感器数据之间进行迁移学习，避免重复采集数据。通过在仿真环境中生成的RGB图像、深度图像、IMU 数据等，模型能够学会在不同传感器输入下进行任务操作。这种迁移机制减少了实际传感器配置下的数据采集成本，并且通过多模态数据的融合提升了模型的性能。

6.4 数据资源提交方式

6.4.1 数据文件标识

数据(资源)文件应通过文件名称来进行标识,其命名规则为:

- a) 文件名称=文件名+文件扩展名;
- b) 命名通常不要含有中文字符和不合法字符等;
- c) 在后续使用过程中不能对数据集进行重命名,否则会造成数据无法回溯的问题,导致数据丢失。

6.4.2 数据资源的提交方式

本文件数据资源提交方式宜满足T/SAIAS 015—2024中的数据资源提交方式。数据资源的提交方式在实体存储介质(如硬盘)、云盘、直连在线、数据空间这四种数据资源提交方式中选定一种方式来实现具体的数字资源交付。

6.4.2.1 实体存储介质

实体存储介质方式是指将数据资源文件按一定的格式和组织形式(如压缩)存入实体存储介质后进行的数据交换方式。本文件附录B提供了这种提交方式的一种参考实现。

6.4.2.2 云盘传输

云盘传输方式是指将数据资源文件按一定的格式和组织形式(如压缩)后通过公有或私有云盘转储所实施的数据交换方式。本文件附录C提供了这种提交方式的一种参考实现。

6.4.2.3 直连在线

直连在线方式是指数据资源供给和接收双方通过光纤专线点对点进行数据传输。这一方式具有较高的安全性和可靠性。若选择该传输方式,所需带宽宜按以下公式进行计算:

文本数据(集)所占总文件长度(MB)*8*3600/拟完成传输的时间(s)=所需带宽(Mb/s)

本文件对采用“直连在线”方式所涉及的数据加密方式不作具体规定。

6.4.2.4 数据空间

数据空间是互相信任的合作伙伴之间的数据关系,每一方都对其数据的存储和共享适用相同的标准和规则。在数据空间中,数据不是集中存储,而是存储数据来源,因此只有在必要时才会(通过语义互操作性)共享数据。

数据空间是数据资产化的有效工具体系。

7 具身智能语料生产

7.1 采集数据完整性检查

在用于语料生产的所有资源数据存入资源数据库前,宜按以下过程检查其格式的统一性和数据的完整性:

- a) 步骤一:所有资源数据以 6.3.1 中所规定各数据表征模式的文件格式之一的形式存在。如承载资源数据的文件不是对应数据表征模式的规定格式之一的,需采取合适的措施实现资源数据文件格式的转换;
- b) 步骤二:对每一个文件进行一次打开操作,在确认打开成功后再执行下一步;否则,将抛弃该文件;
- c) 步骤三:将资源数据以对象方式存入资源数据库。

注:资源数据的统一性和完整性检查的重要性在于,能避免因数据解码失败而导致训练崩溃现象的出现。

7.2 数据资源清洗过程

7.2.1 主要流程

具身智能数据资源清洗过程的主要流程包括数据规范管理、数据整理分类、数据脱敏、数据去噪、数据去重、数据存储与备份等。

具身智能语料数据资源包括现实世界采集的真实数据、合成的虚拟数据，以及二者相结合的虚实数据。

7.2.2 数据规范管理

数据规范管理操作宜包含以下步骤：

- a) 步骤一：统一命名，所有的数据资源需根据 6.3.1 中所规定的文件标识进行统一命名；
- b) 步骤二：统一格式，所有资源数据以 6.3.1 中所规定各数据表征模式的文件格式之一的形式存在。

7.2.3 数据整理分类

数据整理分类主要是对所有资源数据以 6.3.1 中所规定各数据表征模式进行整理和分类。

7.2.4 数据脱密脱敏

数据脱敏主要是对所有数据资源进行脱密脱敏处理，宜包含以下步骤：

- a) 步骤一：将关键帧发到 NSFW 模型中，模型会返回 0 或 1：
0=NOT EQUAL NSFW
1=NSFW
- b) 步骤二：将返回值记录至对应关键帧文本中，若标记为 0 则通过，若标记为 1 则进入人工复核。

7.2.5 数据去噪

数据去噪主要是对所有数据资源进行去噪处理，将文件损坏、图像不完整、畸变过大且无法处理等无法使用的数据去除。

7.2.6 数据去重

数据去重主要是对数据进行查重处理，将重复的数据删除。

7.2.7 数据存储与备份

数据存储与备份主要是数据整理完毕后及时对数据进行存储与备份，遵循一式多份且多个地方存储。

7.3 数据资源标注过程

7.3.1 标注任务制定

任务制定主要是制定具身智能数据资源标注，包含任务解决的问题，重点标注的属性信息，以及标注的注意事项等。

7.3.2 数据标注基本原则

在具身智能使用数据资源进行标注时，宜按照以下过程检查数据是否遵循基本原则：

- a) 步骤一：所有数据资源以时间同步和空间同步进行数据同步，不同传感器（包括摄像头、激光雷达、IMU）的数据在时间上是同步的以及不同传感器的数据在空间坐标系中是对齐的；
- b) 步骤二：目标框宜以矩形框表示；
- c) 步骤三：标注类别以标准分类体系统一，包括物品、车辆、行人、标志等；
- d) 步骤四：标注属性根据不同的标注类别进行定义，包括物品的颜色、行人的前进方向等。

7.3.3 定义坐标系

具身智能中坐标系定义为四种，分别为本体坐标系、相机坐标系、局部坐标系和雷达坐标系，场用定义和作用如表16所示：

表 16 具身智能训练中定义的坐标系类型

序号	坐标系类型	原点位置	X轴方向	Y轴方向	Z轴方向	作用
1	本体坐标系	本体后轴中心或前轴中心	本体前方	本体右侧	指向天空	主要用于获取本体运动状态、相对位置
2	相机坐标系	图像中心或相机镜头中心	图像右侧	图像下方	指向相机内部	主要用于2D图像处理、目标检测
3	局部坐标系	定义的固定点，场景设备	北方	东方	指向天空	主要用于本体导航、路径规划
4	雷达坐标系	图像中心或雷达镜头中心	图像右侧	图像下方	指向雷达内部	主要用于三维空间信息收集

7.3.4 标注方法

7.3.4.1 关键点标注

关键点标注是指将需要标注的元素按照需求位置进行点位标识，从而实现关键点的识别，并由若干关键点的集合形成具体应用所需的语义功能标识。

7.3.4.2 标注框标注

标注框标注是一种对目标对象进行标注的简单处理方式。在标注框标注的帮助下，使模型通过训练能够检测或识别出所需的对象。

框标注包括矩形标注、自由矩形标注、3D框标注，四边形标注和不规则框标注等。

示例：在具身智能训练中，通过标注框将物品的位置确定下来再进行后续的动作操作。

7.3.4.3 区域标注

区域标注是指将图像分成各具特性的区域并提取出感兴趣部分的过程。区域标注包括开区域标注和闭区间标注。

区域标注同时满足均匀性和连通性的条件，其中均匀性指的是该区域中的所有像素点都满足灰度、纹理、彩色等特征的某种相似性准则；连通性是指在该区域内存在的链接任意两点的路径。

与矩形框标注相比，其区域标注更加精确，标注边缘可以是多边形甚至是柔性的。

7.3.4.4 属性标注

属性标注就是用一个或多个标签标注目标对象的属性。

注：对具体应用而言，可在标注文件的元数据部分来申明所用的这些标签，这些标签必须是一个封闭的集合。

7.3.4.5 关键帧标注

关键帧标注就是对视频、点云数据中具有关键动作意义的帧进行属性、区域等内容的文本标注。

8 数据安全

8.1 安全原则

语料数据的建设方宜在组织、制度、人员、平台等方面实施安全策略、执行覆盖资源数据和语料产品全生命周期的安全控制措施。

8.2 安全性评价

语料数据的数据采集、加工、测试和提供的全过程应进行安全性评价，考虑的因素包括：

- a) 宜采取必要和合适的方法确保所涉及数据，不涉及各类秘密、隐私等敏感或违法的信息，未受到蓄意伪造、篡改等行为造成的数据污染，仅暴露在合同或其它强制性规定所明确界定的范围内；
- b) 涉及个人信息的，执行并满足GB/T 35273-2020的相关规定；
- c) 执行并满足《数据安全风险评估方法》、YD/T 4245-2023等的相关规定。

8.3 制度

语料数据的数据采集、加工、测试和提供的全过程考虑建立的制度包括：

a) 数据管理规范：其内容应覆盖上述语料生产的全过程，明确针对数据准备、数据处理、数据交付等不同阶段的数据存储空间和访问控制，并载明不定期组织内部审查的方案；

b) 访问控制及权限管理制度：实施物理的身份验证和授权机制，建立完善的监控和审计机制。

8.4 技术

语料数据的数据采集、存储、加工、测试和提供的全过程所涉及信息系统考虑的软硬件组件包括：

a) 数据网关；

b) 数据加密与备份装置；

c) 安全防范和监控系统；

d) 私域数据存储与传输系统；

e) 数据资产管理工具；

f) 为满足GB/T 22239-2019所需的其它设备或系统。

8.5 人员

参与语料数据的数据采集、加工、测试和提供的任何人员都须签署与其职责相对应的保密协议；这些人员所属机构须建立独立、专业的信息安全团队。

附录 A

(资料性)

具身智能大模型方案

A.1 语言大模型

A.1.1 无监督预训练

无监督预训练主要是利用大规模的文本数据，通过自监督的方式让模型学习文本的结构、语义和上下文关系。最常见的训练方式是使用自回归或自编码模型，包括GPT（生成式预训练）和BERT（双向编码器）。

- a) 典型任务：语言建模（LM）、掩码语言建模（MLM）。
- b) 数据类型：一般为通用文本数据，包括网页、书籍、新闻等。
- c) 常用公开数据集
 - 1) Common Crawl：包含大量从互联网上爬取的网页文本数据，被许多大模型预训练使用，包括 GPT-3 等。
 - 2) Wikipedia：广泛用于多种自然语言处理（NLP）任务的高质量文本来源。
 - 3) BooksCorpus：包含来自在线图书的数据，提供更丰富的上下文信息。
 - 4) OpenWebText：一个试图复制 OpenAI 的 WebText 数据集的开源项目，来源于 Reddit 高评分的链接。
 - 5) The Pile：EleutherAI 组织收集的大型开放数据集，包含 15 个子集，包括 PubMed、GitHub 代码库、FreeLaw 法律文本、YouTube 字幕等。

A.1.2 有监督微调

SFT是在无监督预训练的基础上，使用标注数据集对模型进行微调，进一步提升模型在特定任务上的性能。通常通过少量的任务相关标注数据对模型进行针对性的优化，使其在特定任务上表现更好。

- a) 典型任务：指的是问答、情感分析、对话生成等。
- b) 数据类型：带有标签的结构化数据，包括问答对、对话数据、特定领域的数据等。
- c) 常用公开数据集
 - 1) MS MARCO：微软发布的问答数据集，适用于检索式问答和对话。
 - 2) SQuAD (Stanford Question Answering Dataset)：用于问答系统的高质量数据集，包含大量的问题和答案对。
 - 3) GLUE：包含多个自然语言理解任务的数据集，包括文本分类、句子相似性等。
 - 4) OpenAI's WebGPT：基于网络搜索结果的问答数据集。
 - 5) SuperGLUE：GLUE 的升级版，包含更具挑战性的 NLP 任务。
 - 6) Dialog Datasets (e.g., DailyDialog, Persona-Chat)：用于对话生成的标注数据集，包括对话语境和应答。

A.2 视觉语言大模型

A.2.1 概述

视觉大语言模型通过整合视觉和语言模态数据，实现图像和文本的联合理解和生成。VLM的训练也通常分为无监督预训练和有监督微调（SFT）两个阶段。

A.2.2 无监督预训练

无监督预训练阶段主要通过自监督方法，使用大量图像-文本对数据，让模型学习视觉和语言之间的关联。包括图像-文本对齐、区域掩码预测、视觉词汇建模等。

- a) 典型任务：图像-文本对齐（包括CLIP的对比学习）、图像掩码预测（包括MAE、SimMIM）。
- b) 数据类型：大规模的图像-文本对数据集。
- c) 常用公开数据集：
 - 1) Conceptual Captions：包含 300 万张图像及其相应的自然语言描述，数据来源于网页，广泛用于 VLM 的预训练。

- 2) COCO Captions: Microsoft COCO 数据集的一个版本, 包含了 33 万张图像及其描述, 每张图像有 5 条自然语言描述。
- 3) Visual Genome: 包含 10 万张图像, 每张图像中包含详细的物体和场景关系信息, 有助于模型理解图像内部关系。
- 4) Flickr30k: 包含 3 万张图像, 每张图像都有多个描述性标注, 适用于图像-文本匹配任务。
- 5) LAION-400M & LAION-5B: 由 LAION 团队整理的大规模开放数据集, 包含 4 亿或 50 亿图像-文本对, 用于 CLIP、Stable Diffusion 等模型的预训练。
- 6) YFCC100M: 包含来自 Flickr 的 1 亿张图片和视频, 其中许多带有文本描述。
- 7) RedCaps: 包含 1200 万图像-文本对, 来源于 Reddit, 数据涵盖多个领域, 包含丰富的上下文信息。

A.2.3 有监督微调

在无监督预训练的基础上, VLMs 通常需要通过有监督数据进行微调, 以适应特定任务, 包括图像描述生成、视觉问答、图文检索等。这些任务的目标是提高模型在特定应用领域的表现。

- a) 典型任务: 视觉问答 (VQA)、图像字幕生成、图文匹配、图像标注。
- b) 数据类型: 标注的任务数据集, 包括图像-问答对、图像-描述对等。
- c) 常用公开数据集
 - 1) VQA (Visual Question Answering): 一个用于视觉问答任务的数据集, 包含图像及其相应的问题和答案对。
 - 2) ImageNet (分类任务): 虽然是图像分类数据集, 但在一些模型中用于视觉特征提取或图像-文本匹配。
 - 3) SBU Captions: 包含约 100 万图像和对应的自然语言描述, 适用于图像描述生成任务。
 - 4) NoCaps: 扩展了 COCO 数据集中的描述方式, 使模型能够生成更丰富的图像描述。
 - 5) RefCOCO, RefCOCO+ and RefCOCOg: 用于图像中的目标定位, 通过给定的语言描述定位图像中的特定对象。
 - 6) GQA: 生成性视觉问答数据集, 包含约 113 万个视觉问答对, 测试模型的视觉推理能力。
 - 7) OK-VQA: 一种开放域的视觉问答数据集, 包含更多的现实世界背景问题, 考察模型的常识推理能力。
 - 8) Flickr30k Entities: 在 Flickr30k 的基础上, 将图像中的各个实体进行了详细标注, 用于图像-文本区域对齐任务。图像/视频类数据, 作物体检测或图像分割任务。
 - 9) Pascal VOC: Pascal VOC 是经典的物体检测和分割数据集, 用于图像分类、物体检测、语义分割、实例分割等任务。

A.3 视觉语言动作大模型

视觉语言动作大模型 (VLA) 是指在视觉输入中识别物体并生成相应的动作序列, 或在接收到语言指令后执行指定任务。

微调数据通常包括详细标注的动作轨迹、目标物体位置和环境信息。

- a) 典型任务: 目标物体操控、路径规划、动态环境中的任务执行、视觉问答等。
- b) 数据类型: 标注的任务数据集, 包括物体标注、动作轨迹、指令匹配等。
- c) 常用公开数据集
 - 1) AI2-THOR: 虚拟环境中的视觉-语言任务, 包括目标物体导航和操控任务, 用于机器人任务微调。
 - 2) Epic-Kitchens: 涵盖厨房场景中的详细动作标注和物体操作信息, 用于机器人在复杂环境中的动作识别和执行。
 - 3) R2R (Room-to-Room Navigation): 包括路径规划和导航任务的数据集, 适用于机器人模型的导航和任务执行。
 - 4) RLBench: 提供多种机器人任务操作的高精度数据, 涵盖抓取、放置、旋转等任务, 有助于模型学习精确动作生成。

- 5) ManipulaTHOR: 由 AI2 提供的虚拟机器人操控数据集, 支持机器人在虚拟室内环境中进行操控任务。

国家标准
团体标准
行业标准
地方标准
国际标准

参 考 文 献

- [1] 中国分类主题词表（第二版）
 - [2] 中国图书馆分类法
 - [3] 中国档案分类法
 - [4] GA/T 1717.3-2020 信息安全技术 网络安全事件通报预警 第3部分：数据分类编码与标记标签
技术体系技术规范
 - [5] SF/T 0153-2023 图片真实性鉴定技术规范
 - [6] YD/T 3470-2019 面向公有云服务的文件数据安全标记规范
-