ICS 35.240
CCS L70

# T/SAIAS

## 上海市人工智能行业协会团体标准

T/SAIAS XXX—2025

# 具身智能数据评测方法

Embodied intelligent data evaluation methods

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

上海市人工智能行业协会　　发 布

# 目　　次

# 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：

本文件主要起草人：

本标准首次制定。

首期执行单位：

# 引　言

在具身智能（Embodied Intelligence）研究与应用中，数据质量的优劣直接关系到系统感知、决策和执行效果的可靠性与安全性。本方法旨在提供一套系统、详尽且具有可操作性的评测流程与指标体系，为多模态感知、操作轨迹与交互数据等各类具身智能数据的质量管控和持续改进提供规范与依据。

# 具身智能数据评测方法

## 1 范围

本文件规定了具身智能数据的评测指标、评测流程和评测报告。

本文件适用于具身智能数据的质量评测活动。

## 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。

T/SAIAS 017-2024 人形机器人 分类分级应用指南

T/SAIAS 027-2025 人形机器人 数据集质量评价

## 3 术语和定义

下列术语和定义适用于本文件。

3.1

**具身智能 embodied artificial intelligence**

指通过具身本体与环境交互，能进行环境感知、信息认知、自主决策和采取行动，并能够从经验反馈中实现智能增长和行动自适应的智能系统。

3.2

**具身智能数据 embodied artificial intelligence data**

包含感知（视觉、听觉、触觉）、控制（轨迹、力/力矩）与交互（指令、反馈）等多模态信息的原始与派生数据集，包括在真实环境中采集或在虚拟环境中渲染的数据。

3.3

**多模态对齐 multimodal alignment**

不同模态数据在时间、空间和语义层面的一致性匹配程度。

3.4

**操作轨迹 manipulation trajectory**

记录具身智能系统在执行任务过程中，各执行机构（如机械臂）的位置、姿态与力控信息随时间的变化序列。

3.5

**视觉语言大模型 vision language model (VLM)**

把一张或多张图像+文本问句作为输入，输出文本回答，在具身中充当“看得懂场景并回答高层指令”的机器人大脑，主要用于具身语义理解任务。

3.6

**视觉语言动作大模型 vision language action model (VLA)**

在VLM基础上把文本回答换成可直接执行的低维动作向量或关节角，实现从自然语言到机械臂位姿的端到端机器人控制，主要用于具身操作任务。

3.7

**视觉语言导航大模型 vision language navigation model (VLN)**

在VLM基础上把文本回答换成下一步航点坐标或路径规划信息，让机器人按文本指令走到指定地点，主要用于具身导航任务。

3.8

**数据评测 data evaluation**

应用标准化流程和指标，对数据质量进行定量或定性分析，并形成评测报告的全过程。

3.9

**通专融合 specialized generalist**

构建既具有泛化性，又具备专业能力的人工智能技术路径。

## 4 评测指标

### 4.1 指标维度概述

#### 4.1.1 预处理与校验

对具身智能数据进行标准化处理、完整性校验及安全合规管控，包括统一数据格式与单位、检测并处理缺失值与异常值、实施隐私脱敏并严格校验访问控制，确保数据质量与合规性。

#### 4.1.2 通用质量评测

具身智能数据通用质量评测包含以下内容：

a) 完整性：字段齐全率、缺失率；
b) 一致性：跨模态字段内容对齐率；
c) 真实性：人类、物体、背景自然程度；
d) 易用性：文档说明书、数据读取工具支持；
e) 实用性：可直接调用、数据封装格式兼容各个平台；
f) 扩展性：在不修改文件格式前提下添加新字段能力；
g) 挑战性：数据集的整体难度适中，具体任务包含从简单到困难的梯度划分。

#### 4.1.3 专用领域评测

具身智能数据专用质量评测包含以下内容：

a) 图像：是否包含不同的主体物、背景、环境、以及视角；
b) 文本：是否包含不同的操作、导航、理解任务；
c) 底层视觉：亮度、色度、纹理、对比度、空间信息分布的多样性。

注：在像素层面，底层视觉指标的分布应广泛，以保证各个专用领域上的泛化性。

### 4.2 通用质量维度

具身智能数据评测通用质量维度，应对数据集的每条样本依次进行评测，赋予从0到1之间的分数。在a)~f)中，根据所有样本的平均分数期望，作为数据集在该维度上的得分；在g)中，根据所有样本分数的方差，来量度其分布是否均匀。具体维度如下：

a) 完整性：使用 SQL 工具，仅保留读取成功的字段，检测字段齐全与缺失的比例。根据成功读取的字段数占当前样本总字段数的比例，作为该样本的分数；
b) 一致性：使用多模态大模型，分别读取样本的图像/文本模态字段并进行多模态对齐处理，当图像/视频与文本完全符合时计 1 分，图文存在部分不符物体时计 0.5 分，完全不符计 0 分；

示例：“Evaluate whether the image [{字段 1}] completely matches the text label [{字段 2}].
Please rate score 1 if the data completely or almost completely matches the ground truth on completeness, accuracy, and relevance.
Please rate score 0.5 if the data partly matches the ground truth on completeness, accuracy, and relevance.
Please rate score 0 if the data doesn't match the ground truth on completeness, accuracy, and relevance at all. Please only provide the result in the following format: Score:”

c) 真实性：使用图像或视频质量评价套件对数据集图像进行打分，保证人类、物体、背景自然，没有模糊、抖动、噪声；使用文本质量评价套件对数据集文本进行打分，保证文本通顺，没有拼写、语法、事实错误。将两者的均分作为样本真实性分数；

注： 该维度的图像/视频部分仅适用于仿真数据，真机数据只需计算文本部分。

d) 易用性：整个数据集应提供文档说明书、单条数据应支持主流的数据读取工具，其中图像或视频读取应支持 OpenCV、Pillow、Matplotlib、PyTorch、FFmpeg 的内置读取功能；文本读取应支持 python 基础读取、Panda 函数 read_csv、Numpy 函数 loadtxt、Sk-Learn 函数

load_files、以及 NLTK 的内置读取功能。根据支持的读取工具占所有工具的比例，为样本赋分；

e) 实用性：可直接在多个平台上调用、仿真数据的封装格式兼容 Open X-Embodiment、Mujoco、Robosuite 等主流软件，真机数据的封装格式兼容 TurtleBot3、 Unitree G1、UR5、Robotiq 2F-85、Franka Emika Panda 等主流硬件。根据兼容平台数占总平台数的比例，为样本赋分；

f) 扩展性：在不修改样本文件格式前提下、若可直接以 json 格式添加新字段，则计 1 分；若可通过专有工具添加字段，计 0.5 分；若完全封装则计 0 分；

g) 挑战性：根据 VLA 模型对数据进行推理，根据操作轨迹的成功与否对每条数据的挑战性进行赋分，使用卡方分布计算挑战性梯度，作为整个数据集的 $S_{挑战性}$ 分数：

$$S_{挑战性} = \sum (x - E(x))^2 / E(x) \quad \cdots\cdots (1)$$

式中：

$S_{挑战性}$——调整性分数值；

$x$ ——每道题目推理正确的模型个数；

$E(x)$ ——期望值；

具体测试中，VLA应锚定当前评测榜单上最先进的模型。真机数据应遵照T/SAIAS 017-2024的上肢操作与下肢运动设置，仿真数据应在Libero环境中执行。

**注：** 目前锚定的模型为 1.0 版本（包含 OpenVLA，CogACT，Pi0-Droid，Pi0-Libero，Pi0.5），截至 2025 年 9 月。锚定模型版本将以半年为单位更新，在评测报告中，将注明模型版本号。

### 4.3 专用领域维度

具身智能数据应涵盖不同的主体物、背景、环境、以及视角，从而适应各个专用领域的需求。在数据集中，每个样本的图像或视频字段应包括但不限于：

a) 主体物：应包含日常，电子，机械，工具；

b) 背景：应包含家庭，工业，街道，野外，实验室；

c) 环境：应包含真实光照，二维场景，三维渲染；

d) 视角：应包含第一人称，第三人称。

以下类别均由多模态大模型完成判别，输入图像或视频，输出类别标签，每条样本均包含四个视觉类别标签。每个样本的文本注释字段应包括以下内容：

e) VLA操作：应包含Insert, Move, Pick, Place, Press, Pull, Push, Twist任务；

f) VLN导航：应包含Planning, Detection, Segmentation, Avoidance, Decision Making任务；

g) VLM语义理解：应包含多项选择，开放问答，场景描述任务。

**示例：** 对于专用于某个领域的数据集，可以在以上专用场景中，选择部分进行屏蔽。如专用于“第三人称的操作任务数据集”，只需对(a,b,c,e)维度进行测试。

类似的，任务类别均由多模态大模型完成判别，输入文本，输出类别标签，每条样本均包含三个文本类别标签，最后使用变异系数计算分布是否均衡。例如，主体物共有 4 种类别，则分数遵循$S_{主体物}$：

$$S_{主体物} = 1 - \sum_{T \in (日常，电子，机械，工具)} (n_T - \frac{n}{4})^2 / (\frac{n}{4}) \quad \cdots\cdots (2)$$

式中：

$n_T$——类别T的数据条数；

$n$ ——数据总条数；

像素应基于 OpenCV 图像处理模块，计算亮度、色度、纹理、对比度、空间信息这五个基础属性。其中纹理使用 Laplacian 算子，空间信息使用图像熵。按照 4.2 g)的挑战性公式，将基础属性代入为 $x$，计算其分布的均匀程度。

## 5 评测流程

### 5.1 评测准备

#### 5.1.1 评测目标与范围确认

具身智能数据评测目标与范围应包括以下内容：

a) 将通专融合拆解为通用与专用任务，分别制定数据质量门槛和验收标准；
b) 明确通用数据质量标准（如模糊、抖动、时间一致性）；
c) 明确专用目标场景与对应任务（如抓取、导航、协同操作）。

#### 5.1.2 环境与工具搭建

环境与工具搭建应包括以下内容：

a) 配置统一硬件（相机、力/扭矩传感器）与软件平台；
b) 建立基准数据集与对照实验组。

#### 5.1.3 团队与职责划分

团队与职责划分应包括以下内容：

a) 指定客观评测算法开发人员与参考软件；
b) 明确质量控制与风险应急流程。

### 5.2 评测计划制定

#### 5.2.1 通用指标体系

通用指标体系包括以下内容：

a) 划分通用质量维度，包括完整性、一致性、真实性、易用性、实用性、扩展性、挑战性；
a) 基于项目需求分配指标权重（例如环境适应性 30%、多模态对齐 20%）。

#### 5.2.2 专用领域划分

专用领域划分包括以下内容：

a) 在高层语义上，应划分数据集涵盖的具体细分领域，包括主体物、环境、视角、任务；
b) 在底层视觉上，应定义像素级基本属性的分布范围，包括亮度、色度、饱和度。

#### 5.2.3 采样与实验设计

采样与试验设计应包括以下内容：

a) 确定采样方法（随机、分层、多场景覆盖）；
b) 安排基准任务与对比试验。

#### 5.2.4 风险评估与应对

风险评估与应对包括以下内容：

a) 例举可能存在的风险，如数据缺失、测量漂移、安全隐患；
b) 制定缓解策略与紧急恢复预案。

### 5.3 结果分析与评估

#### 5.3.1 定性评估

定性评估包括以下内容：

a) 根因排查：异常聚类、标签漂移；
b) 在预处理阶段未出现异常。

#### 5.3.2 定量分析

定量分析包括以下内容：

a) 指标计算：实际项数和预期项数的比例，占比越高性能越好；
b) 指标分布：随机选取两个样本计算其差异，期望值越大性能越好；
c) 趋势与对比：相比上一季度基准的变化（柱状图/折线图）。

### 5.4 持续监控与优化

持续监控与优化应包括以下内容：

a） 建立数据质量监控仪表盘与告警；

b） 周期性复评与对比（每季度/每迭代）；

c） 根据用户反馈与模型表现，迭代更新评测指标与方法。

## 6 评测报告

具身智能数据评测应给出评测报告，评测报告内容可参考附录A。

# 附 录 A

（资料性）

## 评测报告

### A.1 评测报告结构

评测报告应包含以下内容：

a） 报告结构：执行摘要、评测背景、方法与流程、结果与分析、问题与建议、附录；
b） 结果呈现：量化指标表格、趋势图、对比分析；
c） 问题诊断：典型异常案例解读、根因与改进方案；
d） 行动计划：优先级、责任人、实施时间。

### A.2 评测报告样例

具身智能数据评测报告样例

报告编号：SAIAS-EID-2025-09-001
锚定版本号：V1.0（2025-09版）
数据集名称：XXX-Embodied-Dataset
评测单位：上海市人工智能行业协会
评测日期：2025-09-01 - 2025-09-05

1 执行摘要
数据规模：8 万条仿真轨迹 + 2 万条真机数据
通用质量平均分：0.92 / 1.00
专用领域平均分：0.85 / 1.00

2 评测背景
目的：验证新采集的“开放遥操作 10-bit/s 数据集”是否满足T/SAIAS XXXX-XXXX《具身智能数据评测方法》质量门槛。
范围：真机厨房场景 7 类日常操作 + 仿真扩展 3 类抗干扰场景。

3 方法与流程
评测模式：☐仿真数据☐真机数据 ☑仿真与真机混合数据
预处理：统一 HDF5-1.4 格式，缺失值 0.06 % → 补零。
锚定模型：OpenVLA，CogACT，PiO-Droid，PiO-Libero，PiO.5。

4 评测结果
通用指标结果：

| 维度 | 得分 | 备注 |
|---|---|---|
| 完整性 | 0.98 | |
| 一致性 | 0.95 | |
| 真实性 | 0.79 | 仿真文本 0.90；仿真图像 0.68 |
| 易用性 | 1 | |
| 实用性 | 1 | |
| 扩展性 | 0.9 | 新增“uncertainty”字段需一次写入 |

| | | |
|---|---|---|
| 挑战性 | 0.83 | 卡方得分 12.4（阈值 ≥ 10.8） |

专用指标结果：

| 维度 | 得分 | 备注 |
|---|---|---|
| 主体物 | 0.85 | |
| 背景 | N.A. | 专用于厨房场景，不适用 |
| 环境 | 0.82 | |
| 视角 | 0.81 | |
| 操作任务 | 0.83 | |
| 导航任务 | 0.94 | |
| 理解任务 | N.A. | 不含该任务，不适用 |

趋势图：（略）

5 异常案例

Case-ID Real05421：光照过曝 → 一致性得分 0.5 → 建议二次采集或剔除。

Case-ID Sim00388：文本标签“Push the handle”对应图像为“Pull”动作 → 建议更正标签并追加审核流程。

Case-ID Sim01007至Case-ID Sim01024：一系列图像样本高度相似，文本标签均为相同任务 → 建议重新设置多种主体物，以及增加任务多样性。

6 改进与行动计划

| 优先级 | 行动项 | 责任人 | 截止日期 |
|---|---|---|---|
| P0 | 重采 500 条过曝样本 | 真机数据-采集组 | 2025-10 |
| P1 | 增加“Pull/Push”二义性审核脚本 | 仿真数据-清洗组 | 2025-11 |
| P2 | 发布数据去重后的新版本，并更新 DOI | 项目组 | 2025-12 |

7 结论

XXX-Embodied-Dataset 数据集在 T/SAIAS XXXX-XXXX《具身智能数据评测方法》评测体系下综合得分 0.89，其中通用部分得分0.92，专用部分得分0.85，可初步用于具身智能模型训练与 10-bit/s 速率基准测试。