arXiv:2603.02390v1 [cs.CV] 2 Mar 2026

OpenMarcie：工业环境中多模态动作识别数据集

Hymalai Bello[1†] Lala Ray[1†] JoannaSorysz[1,2] SunghoSuh[3*] Paul Lukowicz[1,2†]
DFKIKaiserslautern [2]RPTUKaiserslautern-Landau [3]高丽大学{Hymalai.Bello, Lala Shakti Swarup.Ray, Joanna.Sorysz, Paul.Lukowicz}@dfki.de sungho suh@korea.ac.kr
[†]同等贡献 *通讯作者

摘要

智能工厂使用高级技术来优化生产并提高效率。为此，工人活动识别能够准确量化绩效指标，从整体上提高效率，同时有助于保障工人安全。据我们所知，*OpenMarcie*是为制造业环境中的人体动作监测而设计的规模最大的多模态数据集。它包含来自可穿戴设备感知模态以及分布在周围的摄像机的数据。该数据集围绕两种实验设置构建，共涉及*36*名参与者。在第一种设置中，十二名参与者在半真实条件下、不遵循固定规程地执行自行车组装与拆卸任务，从而促进发散性且目标导向的问题解决。第二项实验有二十五名志愿者（*24*份有效数据）参与*3D*打印机组装任务，并提供*3D*打印机制造商说明，以指导志愿者获取程序性知识。该设置还包括顺序协作组装，参与者相互评估并纠正彼此的进度，反映真实世界的制造动态。*OpenMarcie*包含超过*37*小时的第一人称视角与第三人称视角、多模态及多位置数据，具有八种不同的数据类型和*200*多个独立信息通道。该数据集在三项人类活动识别任务上进行了基准测试：活动分类、开放词汇描述和跨模态对齐。该数据集和代码可在*OpenMarcie*。

1. 引言

智能工厂的发展取决于将人类智能融入自动化系统，从而推动形成更高效、更具适应性且以人为本的工业范式。这一整合的基本组成部分人类活动识别（HAR）使系统能够理解、支持并优化复杂工业工作流程中的人类动作。通常，视频数据一直是 HAR 的宝贵信息来源。因此，研究人员投入了大量精力，致力于开发创新的基于视觉的方法和基于视频的数据集，以支持相关领域的发展 [8, 10], , 包括事件相机 [35]。表 1展示了工业环境中用于 HAR 的最先进数据集之间的比较。尽管近期取得了进展，现有工业领域 HAR 数据集仍有三大主要局限：（1）缺乏真正融合可穿戴传感器、视觉和音频并以同步方式采集的多模态数据；（2）依赖高度受限的规程驱动任务，无法体现现实工业工作的开放式、程序性特征；（3）人口统计信息多样性或任务复杂度有限。总体而言，大多数数据集聚焦于简短、孤立的动作，未能捕捉制造业人类工作流程中典型的延续性多步骤活动。此外，人类动作本质上是多模态的，融合了感官、认知和运动过程。动作取决于视觉、听觉和触觉信息，而认知和情绪状态则会影响动作、言语和表情 [30]。动作的多样性、情境依赖性以及与环境的交互进一步增加了复杂性，因此，多模态分析对于准确解读至关重要 [3, 5]。为了有效理解和复制人类活动，人工智能和机器人学必须处理多样化的数据源，包括视频、音频和运动传感器。可穿戴及多位置数据源，以及同步的视觉信息（第一人称和第三人称视角），已被证明具有很高的价值 [11, 37, 38]。在工业环境中尤其如此，因为仅依赖视觉的方法可能引发隐私问题，并增加技术泄露的风险 [6]。因此，我们推出了 OpenMarcie（见图 1）。我们的贡献有三方面。首先，我们推出 OpenMarcie，这是一个包含八种传感模态的多模态数据集，

表 1. 工业领域人体动作识别数据集对比表。多动作表示数据集是否支持并发动作标注（例如，边行走边搬运），此判断基于重叠的动词–对象–工具标注。只有 OpenMarcie 将 **可穿戴设备、第一人称视角 + 外部视角多视图视频、多动作标签和完整的工业覆盖范围**集于一体。

数据集	工业领域	可穿戴设备	Ego	Exo	多视图	多动作	规模 / 重点
InHARD [11]	✓	✓	x	✓	✓	x	人机协作装配、RGB+惯性测量单元
LARa [24]	✓	✓	x	✓	x	x	物流、惯性测量单元+动作捕捉
OpenPack [37]	✓	✓	x	✓	x	x	物流，50h+惯性测量单元/物联网
装配101 [32]	✓	x	✓	✓	✓	x	仅视觉，程序化装配
IKEA-ASM [7]	✓	x	x	✓	✓	x	外部RGB-D，家具装配
HA4M [9]	✓	x	x	✓	x	x	外部RGB-D+红外+姿态
HA-ViD [38]	✓	x	x	✓	✓	x	丰富的语义标签、多路线
IndustReal [31]	✓	x	✓	x	x	x	PSR、仅自我视角、误差建模
Ego-Exo4D [17]	（∼**6% 独立**）	✓	✓	✓	✓	x	1,200 小时熟练活动
OpenMarcie（我们的数据集）	**（100%）**✓	✓	✓	✓	✓	✓	工业多任务处理，8 种模态

包括激光雷达深度、多视角姿态估计、目标检测和精确定位。其次，该数据集记录了真实的工业装配和拆卸任务，其中参与者的决策会影响可选路径，从而能够研究程序性知识获取。第三，我们提供了来自 36 名志愿者的丰富元数据、同步的可穿戴传感数据流、语音转文本叙述，以及详细的人工标注，以捕捉重叠的多原子动作，支持对复杂工作流程进行细粒度分析。

2. OpenMarcie数据集

定义了两种主要的实验设置：临时场景和程序性场景，分别涉及自行车和3D打印机的组装/拆卸，如图 1所示。这些场景涵盖了多种技能，包括：机械技能，如工具使用、部件对齐和紧固流程；认知过程，包括解读视觉说明或示意图，以及在出现不一致时解决问题；精细动作技能，这对于精准操作细小或复杂的组件至关重要。我们选择了两项对比鲜明的任务：自行车组装是一项熟悉、目标明确且动作开放式的任务；3D打印机组装则是一项高度程序化的任务，需要解读详细说明并处理陌生部件。两者共同涵盖了非脚本化维修和结构化生产线装配，同时也包括布线、线缆布设、滑轮安装、拆包以及零件整理等更广泛的活动。这种多样性使得我们能够针对真实工业工作流程进行基准测试。目前的基准测试侧重于活动分类、开放词汇图像描述和跨模态对齐，为评估该数据集的范围和实用性建立了坚实的基线。对于两种场景，我们尽可能合理地布置了3台不带偏振片的ZED X AI立体相机，以覆盖整个实验室视野。图 2展示了实验室的设置，以及场景**(a)** 临时自行车组装和**(b)**程序性场景中的3D打印机实验所用外向心相机的示例视图。ZED 立体相机配备了功能丰富的软件开发工具包，并允许将定制的人工智能方法纳入流水线。表2 展示了两种场景中记录的感知模态。可用的原始通道总数约为282个，其中包括各种传感器布置。此外，每位参与者都针对每种场景审阅并签署了伦理协议，以确保符合《赫尔辛基宣言》。德国人工智能研究中心的伦理委员会批准了本研究，协议编号为HRW-35/24和SMD-30/24。

2.1. 临时场景（a）：自行车

自行车的装配与拆卸被提出作为一项结构化、目标导向的任务，参与者可以自主决定如何完成。这反映了许多工业和实验设置中常见的实用型、结果导向场景的本质。常见错误，例如螺栓未正确拧紧或制动器未对准，通常很容易发现并纠正，从而提升了任务的安全性和教学价值。由于自行车广为人知且易于获取，这项任务尤其适用于培训环境、教育环境和研究应用，包括人机协作与任务规划研究。志愿者佩戴了图 3所示的可穿戴传感器套件。该套件包括安装在双手和头部（集成于眼镜框架中）的惯性测量单元（IMU）和气压传感器。此外，胸部和肩部还配备了热成像传感器和光谱仪传感器。系统还配有一台第一人称 RGB-D 摄像机，其中包含基于激光雷达的深度感知。图 3还展示了任务期间采集的多模态传感器数据的示例可视化结果。对于真实标注，我们选取了信息量最大的外中心视角（自行车装配时通常为门侧摄像机；图 2），并由我们（人类）使用一种感知意图的动词–对象–工具方案进行手动标注（图 4）。由此得到与任务意图相吻合、具有语义意义的片段，其中包括多标签情形（例如，边行走边搬运）。每条标注（动词、工具、物体、备注）在

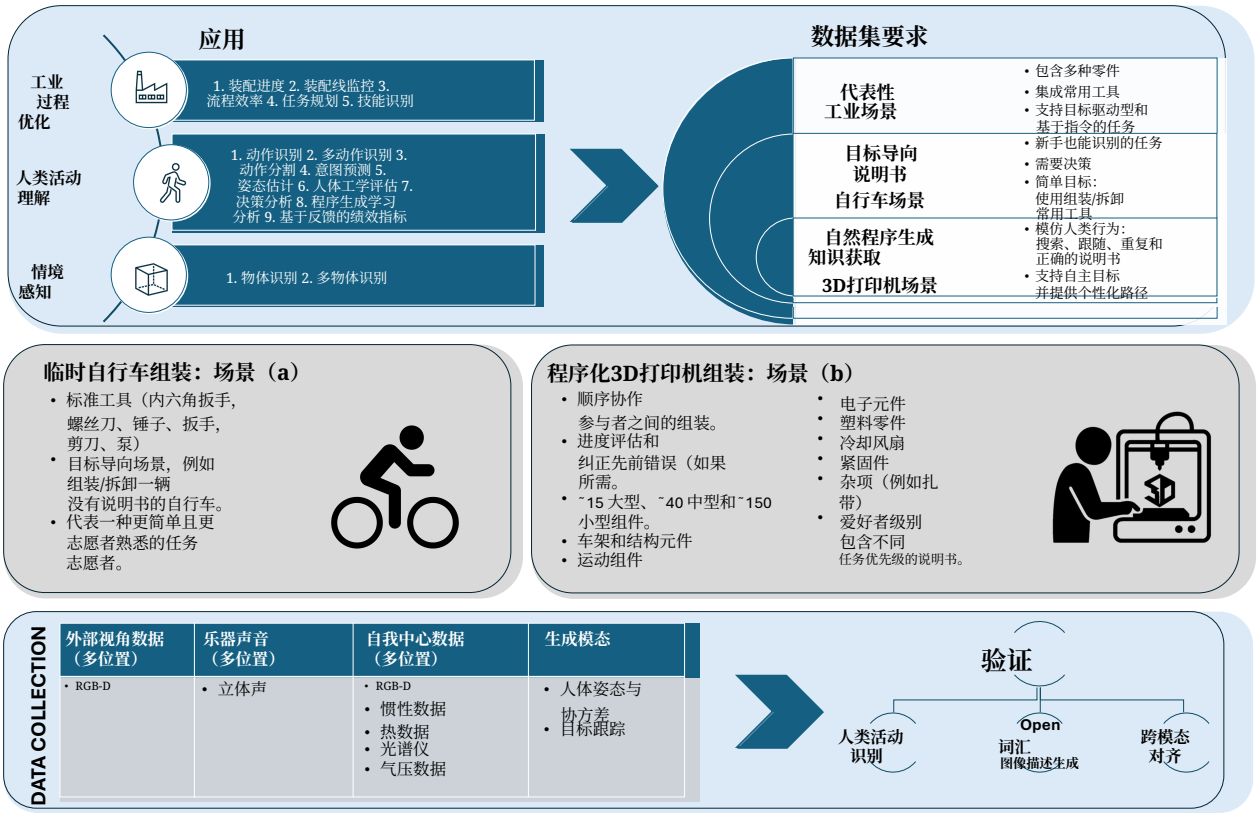


图 1。OpenMarcie 是一个面向工业环境中多模态动作识别的开放数据集。**顶部** 其应用包括工业流程优化、人类活动理解和情境感知。**中部** 两个装配场景分别呈现了临时的目标导向动作和程序生成场景，以促进自然的知识获取。**底部** 采集的多模态数据包含 200 多个独立通道，并使用三个不同的基准测试进行验证。

所有模态（第一人称视频、惯性数据、热成像数据、音频）之间进行时间对齐。我们还使用 GPT-4o 生成简洁的自然语言描述 [1]（参见第 9节），以提供软标签，在保留人工创建的真实标注的同时，提升一致性和可解释性。

2.2. 程序性场景（b）：3D打印机

组装一台3D打印机需要付出大量认知努力，以理解书面和视频说明书。此外，还需要处理螺丝等小型组件，以及金属主框架等较大的硬件部件。因此，这一场景为监测人类动作识别提供了理想环境，与工业装配线上常见的任务非常相似。本场景使用 Assemble Yourself 原装 Prusa i3 MK3S+套件。制造商提供了一套详尽的装配说明书，每个塑料部件都附有其3D模型STL文件，便于重新打印。这也意味着，将3D模型与OpenMarcie的第一人称视频相结合，可借助预先定义的部件词典增强目标检测。这种方法能够直观地监测打印机的装配状态，并引导组装者进行正确的下一步操作，从而加快组装进程

表2. 各场景的感知模态信息

场景	模态	位置	Ch
临时**(a)**	IMU	手腕、前额	21
	磁力计	手腕、前额	9
	气压计	手腕，Head	3
	温度	手腕、前额	3
	光谱仪	胸部，右肩	24
	热成像相机	胸部，右肩	(8×8)×2
	RGB-激光雷达	胸部	4
	立体声	胸部	2
	RGBD	2 个外向心相机	8
程序生成 **(b)**	IMU	手腕	20
	磁力计	手腕	6
	气压计	手腕	2
	温度	手腕，前额，胸部	4
	RGBD-IMU	前额，胸部	28
	RGB-激光雷达	胸部、桌子	8
	立体声	胸部，桌子	4
	RGBD	2 个外向心相机	8
两个场景	17 种模态	9 个位置	282 个通道

流程。实验开始前，每位志愿者都必须观看一段时长 30 分钟的视频，内容是完整的 Prusa 装配指南，可在线访问：https://www.youtube.com/watch?v=uToqSlh64R4。这样可以确保他们理解任务的关键要点，无需观察者进一步解释或提供说明，从而能够独立完成打印机装配。我们的想法是

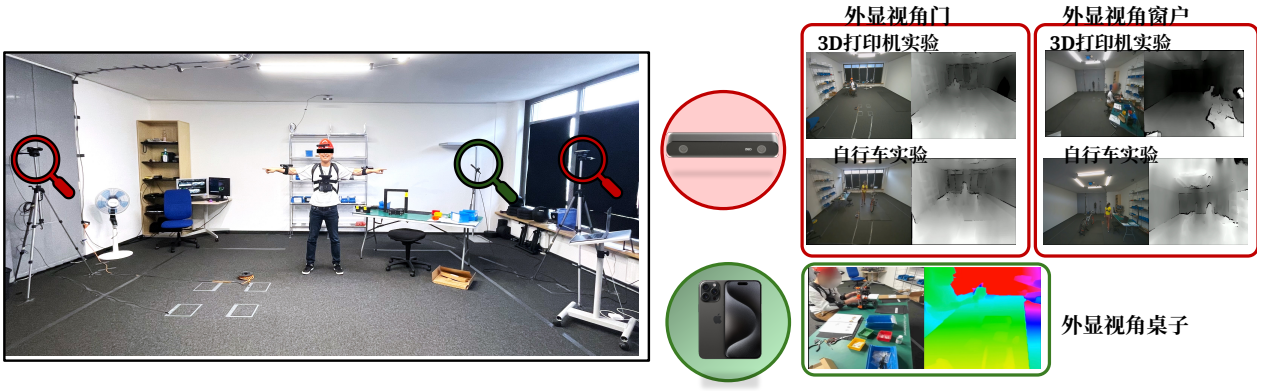


图2. 实验室布置，以及外显式RGBD相机的示例视图。

由一位志愿者仅依据包装盒内的说明和设置开始装配，约有一小时的时间尽可能多地完成装配。随后，下一位志愿者从上一位参与者停下的地方继续，依次重复这一过程。这种安排要求每位后续参与者评估当前的装配状态，了解前一位志愿者的进度，并判断接下来如何进行，从而增加任务的认知挑战。参与者佩戴了4 台 ZED 2 AI立体相机。其中两台 ZED 安装在手腕上（仅用于采集惯性测量单元和气压计信号，立体相机处于禁用状态），一台通过头盔安装在前额，另一台通过绑带安装在胸部，用于捕捉视觉数据流。参与者胸前安装了一部 iPhone 15 Pro，手机保护壳上的摄像头和激光雷达传感器朝向前方。数据通过 USB线缆传输到 Jetson Orin AGX，该设备放在背包内，并与一台 Green Cell 26800mAh 128W充电宝一同使用，以便志愿者连续一小时无缝记录数据。实验期间，一名外部人工观察员实时讲述参与者的动作及场景背景信息。随后，使用 faster-whisper（v1.1.1）和 ctranslate2（v4.4.0），采用“large-v3”模型配置 [28] 对这些讲述的音频录音进行转录。使用这些方法生成的软标签示例见图4 **场景（b）**。随后，通过两阶段流程获得硬标签。该流程包括使用 Deepseek-r 迭代识别并完善离散动作类别1 [18], 然后使用 GPT-4o[1] 通过提示工程将软标签句子转换为适用于模型训练的稳定硬标签。这一流程由人工标注和随后进行的两阶段基于大语言模型的转换组成，最终生成用于下游模型训练的硬标签。在场景（b）中，我们采用了利用转录讲述内容的两阶段流程，这与场景（a）不同。尽管方法有所不同，两者都生成一致的多模态标注格式。为验证大语言模型生成的标签，我们进行双向交叉一致性检查（结构化→图像描述生成→结构化），实现了较强的对齐效果（场景a中的宏 F1 = 0.715；场景b中的 METEOR = 0.531；参见第 9节）。这些结果支持了大语言模型作为我们部分监督流程中结构化翻译器的可靠性。此外，传感器的放置经过调整，以适应各场景特定的人体工学要求和任务需求。尽管这一设计导致了一些差异——尤其是场景（a）（自行车装配）与场景（b）（3D打印机组装）之间——但两项任务中的关键可穿戴模态保持一致。最值得注意的是，惯性测量单元（IMU）佩戴在手腕上，可穿戴激光雷达设备安装在胸部，并且两种场景中均使用立体声麦克风。这些共用的传感器位置有助于进行有意义的多模态比较，尤其适用于人类活动识别和图像描述基准。如3所示，尽管各场景的适应方式有所不同，这种重叠在视觉上仍十分明显，并支持对比分析。

3. 统计数据

研究涉及一组多元化的参与者，其人口统计信息和职业特征总结于图5中。参与者的身高介于150至193厘米之间，其中人数最多的一组（31%）身高在181至193厘米之间。年龄分布偏向年轻群体：47%年龄为22至24岁，25%为25至29岁，28%为30至37岁。参与者自行报告了其装配相关任务的经验水平：57%为初学者，37%为中级，6%为高级。在学术和职业背景方面，工程师占多数（72%），其次是计算机科学家（14%）、生物学家（8%）、物理学家（3%）和管理人员（3%）。这一构成表明，参与者群体以技术从业者为主，职业生涯早期者居多，他们在组装任务方面拥有不同程度的实践经验。参与者来自20多个国家，国际多样性为研究增添了文化和经验层面的多样性。此外，大多数参与者（31人）是

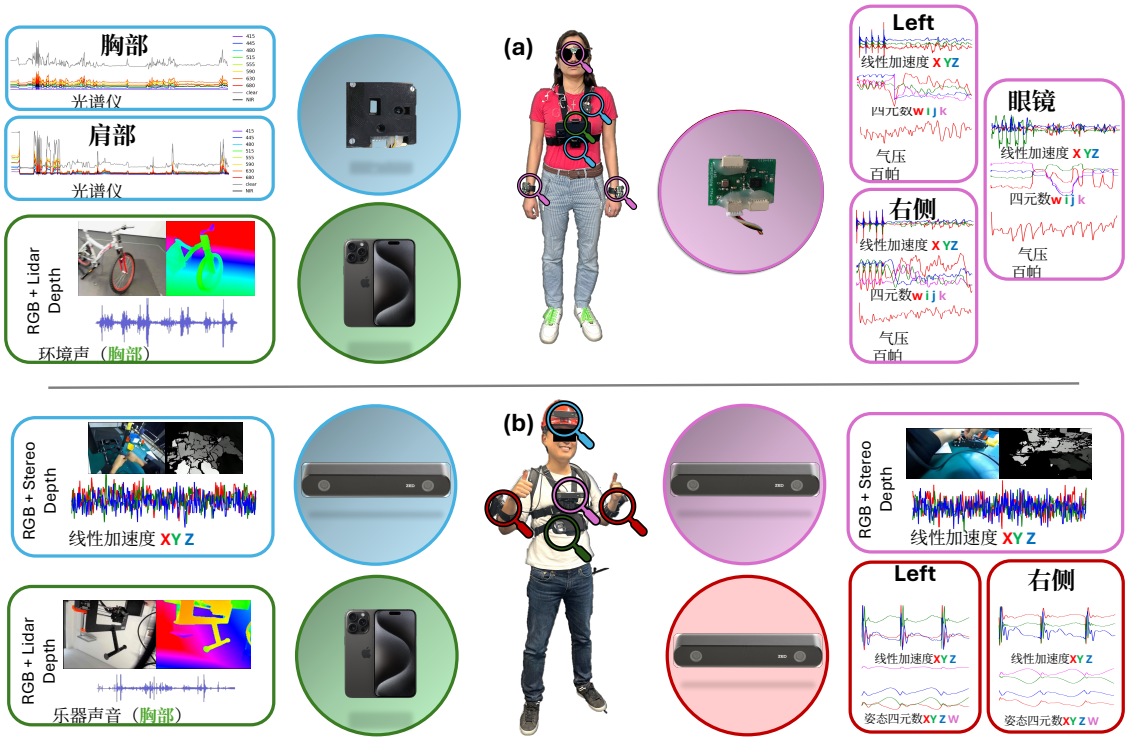


图3. 参与者可穿戴配置与传感器信号。（**a**）临时场景中的参与者：自行车组装。（**b**）程序性场景中的参与者：3D打印机组装。腕部佩戴的ZED仅提供惯性测量单元和气压计数据（立体相机已禁用），而头部与胸部相机采集视觉数据流。传感器布置根据各场景的人体工学与任务需求进行了调整。

右利手，另有少数（5人）自认为是左利手。尽管参与者群体以具有工程背景的右利手为主，可能限制了人口统计层面的普遍性，但这一群体反映了典型的工业从业者，从而增强了生态效度。参与者报告了自行车和3D打印机的所有权及使用习惯（见图6）。多数参与者表示拥有自行车，而拥有3D打印机的只占一小部分。使用模式也反映了这种所有权差异：参与者使用自行车的频率高于3D打印机，使用频率包括每天、每周或每月。相比之下，大多数参与者表示很少使用或从不使用3D打印机。这些差异表明，参与者对自行车更加熟悉，也有更多亲自动手的经验，因此自行车相关任务对他们而言更为直观；而由于此前接触有限，3D打印机任务可能需要更多指导或探索。参与者的特征为解读研究中观察到的任务表现和交互行为提供了宝贵背景。参与者多样的技术背景、不同的装配经验水平以及广泛的国际代表性，共同构成了丰富多元的数据集。此外，与3D打印机相比，参与者对自行车更加熟悉，这为

临时任务与程序性任务之间形成自然对比，进一步说明所选场景对于研究不同先验知识与技能水平下的目标导向活动具有重要意义。图7 展示了OpenMarcie中活动分布的一个示例。这些动作代表人类在工业场景中、尤其是装配任务期间预期执行的一系列多样化活动。该数据集聚焦于两种不同的场景。图7 **左侧**为临时自行车场景的活动分布（**场景（a）**）。最常见的活动是拧紧/拧松和检查，因为这些动作是处理自行车零件不可或缺的组成部分。相比之下，锤击或躺下出现在装配过程中更为明确的阶段。图7 **右侧** 展示了3D打印机组装程序性场景的活动分布。移动物体（“移动”）、调整和阅读是主要类别。与临时场景相比，未标注类在3D打印机情形中不那么突出。

4. 验证基准

为评估 OpenMarcie 的实用性和通用性，我们围绕三项核心任务建立验证基准：人类活动识别（HAR）、开放词汇描述和跨模态对齐。这些任务旨在反映工业和任务驱动环境中的关键挑战，包括——



图4。活动示例的第一人称视角和外部视角，并附有标注（参见第 9 节）。这些示例包括两个场景的软标签和硬标签： **(a)** 临时自行车组装/拆卸任务，以及 **(b)** 程序化3D打印机建造任务。

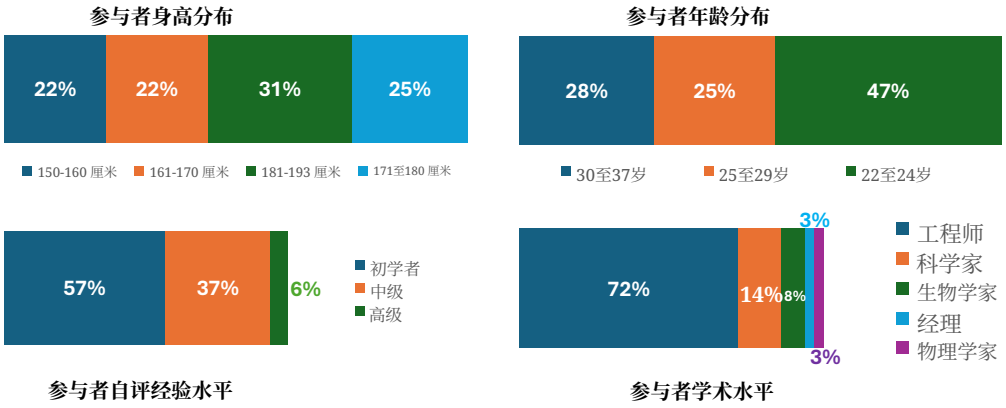


图5. 参与者统计数据。**左上** 身高分布范围为150至193厘米。**右上** 年龄分布范围为22至37岁。**左下** 参与者自我报告的装配任务经验水平，分为初学者、中级和高级。**右下** 参与者的学术水平，其中工程师占多数，管理岗位仅占样本的3%。

人类活动识别支持安全监测、技能评估和机器人模仿等下游应用。开放词汇描述有助于程序生成文档，并支持自然的人机交互。跨——

模态对齐对于传感器替代、检索和迁移至关重要，尤其是在传感模态可能受限或不同步的环境中。这些任务也旨在弥补先前数据集的已知不足，例如动作窗口过短、缺乏语言基础，以及传感器有限——

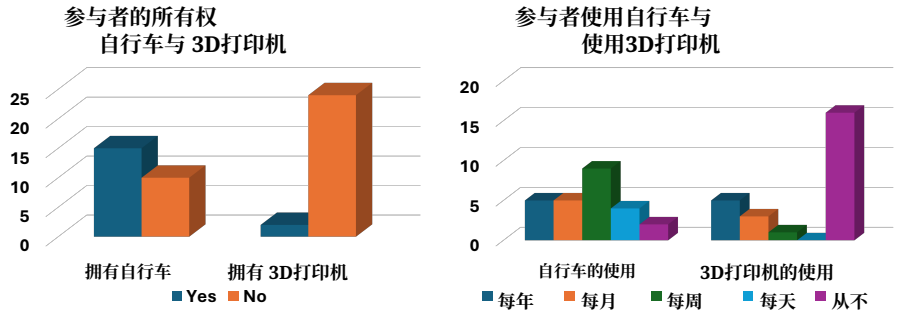


图 6. 参与者对自行车和 3D 打印机的所有权及使用习惯。

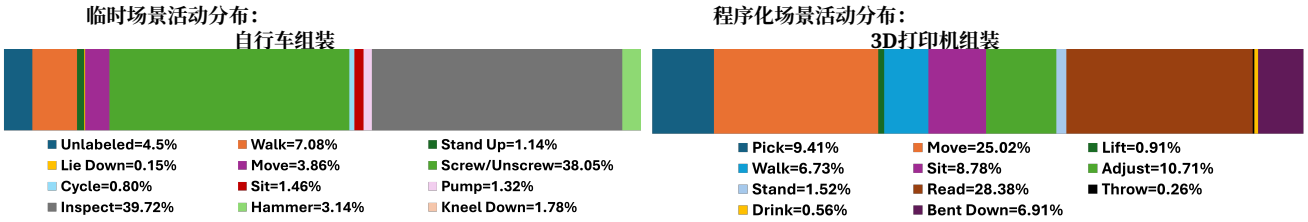


图 7. OpenMarcie 在临时场景和程序性场景中的活动分布。

表 3. 场景（a）：自行车组装和场景（b）：3D打印机组装中，人类活动识别的宏平均F1分数（包含和不包含空类）。

模态	场景（a）		场景（b）	
	宏平均 F1（↑）			
	无空类	Null	无空类	Null
惯性(I)	0.834±0.007	0.811±0.007	0.750±0.015	0.674±0.003
声学(A)	0.489±0.018	0.469±0.017	0.425±0.004	0.432±0.005
视觉(V)	0.757±0.011	0.729±0.011	0.705±0.004	0.655±0.003
I +A	0.803±0.012	0.782±0.010	0.744±0.004	0.666±0.003
A +V	0.739±0.016	0.714±0.013	0.695±0.003	0.646±0.003
I +V	**0.882±0.009**	**0.851±0.009**	**0.773±0.000**	**0.685±0.000**
I +A +V	0.859±0.010	0.831±0.011	0.763±0.003	0.676±0.003

表 4. 场景（a）：自行车装配和场景（b）：3D打印机组装的开放词汇描述结果，包括有无空类时的余弦相似度值。

模态	场景（a）		场景（b）	
	余弦相似度 (↑)			
	无空类	Null	无空类	Null
惯性(I)	0.518±0.023	0.501±0.022	0.642±0.002	0.640±0.002
声学(A)	0.361±0.030	0.341±0.018	0.316±0.003	0.323±0.004
视觉(V)	0.479±0.016	0.463±0.014	0.632±0.002	0.631±0.003
电流 +A	0.512±0.021	0.493±0.020	0.644±0.002	0.641±0.003
电压 +V	0.466±0.025	0.444±0.012	0.626±0.003	0.625±0.004
电流 +V	**0.561±0.016**	**0.531±0.014**	**0.655±0.000**	**0.655±0.000**
I +A +V	0.547±0.020	0.519±0.017	0.647±0.001	0.646±0.003

多样性——定位 OpenMarcie，将其作为面向真实世界多模态智能的测试平台。

工业场景中的人类活动识别颇具挑战性，因为活动具有程序生成性、目标导向性且可能并发发生。

与先前数据集不同 [37] 后者包含简短、脚本化的动作，OpenMarcie 捕获了长时间、未分段的序列，其中包含自然变化、工具使用和多样的动作路径。其丰富的多模态设置配备多位置传感器、同步的自我视角／外部视角以及口头叙述，为以人为中心的自动化提供了一个逼真的测试平台。我们使用三种模态评估人类活动识别（HAR）：第一人称视频、右手惯性测量单元和乐器声音。每种模态均使用专属模型单独测试：用于视频的视觉变换器（ViT）[14]、用于惯性测量单元的 Deep-卷积长短期记忆网络 [33] 以及用于音频的 EnCodec [12]，后者搭配时序分类器，在一个具有受试者不重叠划分的 12 类活动任务上进行评估。对于多模态评估，我们探索了所有两两组合以及使用后期融合Transformer [26] 进行的完整三路融合，该Transformer整合了来自各个数据流的时序编码嵌入。这一设置量化了各模态单独的判别能力及其互补性。表3 展示了两种场景（a、b）下各模态的宏平均F1分数，分别对应纳入和未纳入空类的情况。惯性 + 视觉始终优于所有其他组合，取得了最高的宏平均F1分数。多模态融合通常比单模态输入表现更好，尤其是相较于单独使用声学模态；后者在所有条件下表现最差。

开放词汇图像描述 会生成动作与场景的自由形式描述，这对文档记录、操作员反馈和人机交接等工业任务至关重要。不同于先前聚焦于简短、通用活动并采用模板化标签的数据集 [21], OpenMarcie 提供与视觉和音频流对齐的非脚本化口头叙述。这使得在细粒度、目标导向场景中生成有依据的语言成为可能，并可用于研究多模态上下文，尤其是语言意图，如何提升图像描述性能。

表5. C ross-模态对齐结果：场景（a）：自行车组装与场景（b）：3D打印机组装 .

模态	场景（a）			场景（b）		
	R@1	R@5	Top-1	R@1	R@5	Top-1
惯性 (I) + 文本(T)	0.324±0.016	0.655±0.025	0.481±0.018	0.312±0.016	0.642±0.026	0.468±0.019
声学 (A) +T	0.241±0.014	0.583±0.025	0.342±0.016	0.227±0.013	0.567±0.022	0.329±0.015
视觉 (V) + T	0.437±0.015	0.768±0.017	0.556±0.016	0.421±0.013	0.751±0.018	0.541±0.014
图像 + 音频 + 文本	0.347±0.014	0.679±0.019	0.495±0.017	0.334±0.015	0.663±0.017	0.479±0.018
音频 + 视频 + 文本	0.412±0.013	0.740±0.020	0.533±0.015	0.395±0.014	0.723±0.019	0.517±0.014
图像 + 视频 + 文本	**0.485±0.014**	**0.803±0.019**	**0.587±0.016**	**0.467±0.013**	**0.787±0.015**	**0.570±0.016**
I + A + V + T	0.470±0.015	0.795±0.019	0.579±0.016	0.453±0.014	0.779±0.018	0.563±0.016

与人类活动识别（HAR）类似，我们使用模态专属编码器，回归根据参与者叙述生成的句子嵌入，方法如 OV-HAR 所提出 [29]。图像描述被建模为嵌入预测，输出通过 Vec2Text [23] 使用嵌入检索进行解码。我们评估了单模态、成对模态和融合设置。该方法无需大语言模型即可实现高效的开放词汇图像描述。表 4 报告了图像描述的余弦相似度分数，显示出与人类活动识别相似的趋势。惯性 + 与视觉模态在两个数据集上均取得最佳表现，而单独使用声学模态时表现最差。将声学模态添加到其他模态中仅带来微小提升，甚至没有提升，这表明与惯性和视觉模态相比，其贡献有限。

跨模态对齐旨在学习跨感官模态的共享表征，从而实现检索、识别和迁移。尽管 CLIP 等模型 [27],Multi3Net [15]在网络规模数据上表现出色，但缺乏结构化且传感器丰富的录制数据。OpenMarcie 提供了一个真实的测试平台，包含来自目标驱动任务的同步视频、音频和惯性测量单元数据，支持在超越互联网规模基准测试的复杂现实场景中开展对齐研究。受 ImageBind 启发 [16], 我们采用对比学习，将第一人称视频、惯性测量单元数据、音频和语言对齐到共享嵌入空间中。每种模态都由专用骨干网络编码，并使用针对同步正样本对的多模态 InfoNCE 损失进行训练。我们评估所有 2、3 和 4 种模态组合，从而能够细致分析模态间的对齐情况和共享信息。如表5所示，在两个场景中，将惯性和视觉模态与文本结合可获得最强的对齐性能。视觉和文本单独使用时表现良好；声学特征单独使用时贡献较小，但与其他模态结合后性能有所提升。总体而言，更丰富的模态融合能够改善对齐效果，不过在组合全部模态后，增益趋于平缓。基于嵌入的方法可高效、可扩展地评估各项任务，凸显了 OpenMarcie 在研究真实世界中传感器丰富的人类活动理解方面的价值。在人类活动识别、开放词汇描述和跨模态对齐这所有任务中，多模态融合的表现始终优于单模态输入，其中惯性和视觉模态提供了最强的信号。单独使用音频提供的信息有限，但与其他模态结合时能发挥有意义的作用。此外，由于数据是在测试台而非真实工厂中采集的，因此缺少真实的工业噪声（例如机械设备噪声、振动声）。尽管存在这些限制，该模态在多模态融合中仍然有用，并凸显了在现实条件下开展隐私保护感知的机会。补充材料中提供了更多实验，进一步探究音频性能有限的原因。

5. 结论

OpenMarcie 是一个面向人类活动识别、跨模态学习和工业自动化的大规模多模态数据集。它记录了临时性和程序生成的装配任务，包含同步的第一人称/第三人称视频、可穿戴传感器、匿名化音频和文本叙述，为具身人工智能提供了可靠的基准测试。该数据集通过涉及错误纠正和任务监控的连续协作活动，呈现了真实的工业工作流程。参与者主要是右利手工程师，这限制了人口统计信息方面的泛化能力，但这与典型的工业从业者群体相符，从而增强了生态效度。标注仅涵盖了数据集潜力的一部分，因为其多视角录制支持未来对物体、动作、交互和姿态进行标注。传感器布置根据不同场景进行了人体工学调整，但腕部惯性测量单元、胸部佩戴式激光雷达和立体声麦克风等关键模态保持一致，因此可以进行多模态比较。总体而言，该数据集为工业环境中的细粒度活动识别、人机协作和情境感知人工智能研究提供了灵活的基础。

致谢

本研究得到了德国联邦教育与研究部（BMBF）通过 Cross-Act 项目（资助协议编号：01IW25001）提供的支持。本研究还获得了由 MSIT 资助的 IITP 拨款（RS-2019-II190079，人工智能研究生院项目（高丽大学）和 IITP-2026-RS-2024-00436857，IITP-ITRC），以及由 MOTIR 资助的技术创新计划（RS-2025-25453819）的支持。

参考资料

[1] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。GPT-4 技术报告，arXiv 预印本 arXiv:2303.08774，2023。3、4、7[2] Satanjeev Banerjee 和 Alon Lavie。METEOR：一种用于机器翻译评估的自动指标，与人工判断的相关性更高。载于 ACL 机器翻译和／或摘要的内在与外在评估指标研讨会论文集，第65–72页，2005。6、11[3] Hymalai Bello。用于可穿戴活动识别的单模态与多模态传感器融合。载于2024年IEEE普适计算与通信国际会议研讨会及其他相关活动（PerCom研讨会），第364–365页。IEEE，2024。1[4] Hymalai Bello、Jhonny Rodriguez 和 Paul Lukowicz。基于 RFID同步的可穿戴差分气压计垂直手部位置估计。载于 EAI 身体区域网络国际会议，第 24–33 页。Springer，2019。13[5] Hymalai Bello、Luis Alfredo Sanchez Marin、Sungho Suh、Bo Zhou 和 Paul Lukowicz。InMyFace：用于可穿戴面部活动识别的基于惯性和机械肌动图的传感器融合。Information Fusion，99：101886，2023。1[6] Hymalai Bello、Daniel Geißler、Sungho Suh、Bo Zhou 和 Paul Lukowicz。TSAK：面向制造生产线中可穿戴模态与模型高效优化的两阶段语义感知知识蒸馏。载于国际模式识别会议，第201–216页。Springer，2025。1[7] Yizhak Ben-Shabat、Xin Yu、Fatemeh Saleh、Dylan Campbell、Cristian Rodriguez-Opazo、Hongdong Li 和 Stephen Gould。IKEA ASM数据集：通过动作、物体和姿态理解人们装配家具的过程。载于 IEEE/CVF 冬季计算机视觉应用会议论文集，第847–859页，2021。2、1[8] Fabian Caba Heilbron、Victor Escorcia、Bernard Ghanem 和 Juan Carlos Niebles。ActivityNet：用于人类活动理解的大规模视频基准测试。载于 IEEE计算机视觉与模式识别会议论文集，第961–970页，2015。1[9] Grazia Cicirelli、Roberto Marani、Laura Romeo、Manuel Garc´ıa Dom´ınguez、J´onathan Heras、Anna G Perri 和 Tiziana D’Orazio。HA4M数据集：用于制造业人类动作识别的装配任务多模态监测。Scientific Data，9（745）：1–2，2022。[10] MMAction2贡献者。OpenMMLab 新一代视频理解工具箱与基准测试。https://github.com/open-mmlab/mmaction2，2020。1[11] Mejdi Dallel、Vincent Havard、David Baudry 和 Xavier Savatier。InHARD——工业协作机器人领域的工业人类动作识别数据集。载于2020年IEEE人机系统国际会议（ICHMS），第1–6页。IEEE，2020。1、2

[12] Alexandre D´efossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。《高保真神经音频压缩》。arXiv 预印本 arXiv:2210.13438，2022。7，9[13] Blackmagic Design。DaVinci Resolve。https : / / www . blackmagicdesign . c om / products / davinciresolve/。访问日期：2025-05-08。5[14] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Syl- vain Gelly 等。《一张图像胜过 16×16 个词：用于大规模图像识别的变换器》。arXiv 预印本 arXiv:2010.11929，2020。7，9[15] Vitor Fortes Rey、Lala Shakti Swarup Ray、Qingxin Xia、Kaishun Wu 和 Paul Lukowicz。《通过联合表征语言、姿态和合成惯性测量单元，提升基于惯性手部数据的人类活动识别》。载于《2024 年 ACM 可穿戴计算机国际研讨会论文集》，第 25–31 页，2024 年 8 月。[16] Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。《ImageBind：一个嵌入空间绑定万物》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 15180–15190 页，2023。8，11[17] Kristen Grauman、Andrew Westbury、Lorenzo Torresani、Kris Kitani、Jitendra Malik、Triantafyllos Afouras、Kumar Ashutosh、Vijay Baiyya、Siddhant Bansal、Bikram Boote 等。《Ego-Exo4D：从第一人称和第三人称视角理解熟练的人类活动》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 19383–19400 页，2024。2[18] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。《DeepSeek-R1：通过强化学习激发大语言模型的推理能力》。arXiv 预印本 arXiv:2501.12948，2025。4，7[19] 英特尔。Audacity 的 OpenVINO 人工智能插件：音乐分离。https://github.com/intel/openvino-plugins-ai-audacity。访问日期：2025-05-08。5[20] Andrew Jaegle、Sebastian Borgeaud、Jean-Baptiste Alayrac、Carl Doersch、Catalin Ionescu、David Ding、Skanda Koppula、Daniel Zoran、Andrew Brock、Evan Shelhamer 等。《Perceiver IO：面向结构化输入与输出的通用架构》。arXiv 预印本 arXiv:2107.14795，2021。14[21] Huan Liu、Guangbin Wang、Ting Huang、Ping He、Martin Skitmore 和 Xiaochun Luo。《通过图像描述生成呈现建筑施工活动场景》。Automation in Construction，119:103334，2020。7[22] Seungwhan Moon、Andrea Madotto、Zhaojiang Lin、Tushar Nagarajan、Matt Smith、Shashank Jain、Chun-Fu Yeh、Prakash Murugesan、Peyman Heidari、Yue Liu 等。《Anymal：高效且可扩展的任意模态增强语言模型》。载于《2024 年自然语言处理实证方法会议论文集：产业分会》，第 1314–1332 页，2024。14[23] John X Morris、Volodymyr Kuleshov、Vitaly Shmatikov 和 Alexander M Rush。《文本嵌入揭示的内容几乎与文本本身一样多》。

arXiv 预印本 arXiv:2310.06816，2023。8，10[24] Friedrich Niemann、Christopher Reining、Fernando Moya Rueda、Nilah Ravi Nair、Janine Anika Steffens、Gernot A Fink 和 Michael Ten Hompel。《LARa：利用语义属性构建物流人类活动识别数据集》。Sensors，20(15):4083，2020。2，1[25] ORB-HD。deface。https://github.com/ORB-HD/deface。访问日期：2025-05-08。5[26] Yagya Raj Pandeya 和 Joonwhoan Lee。《基于深度学习的多模态信息晚期融合，用于音乐视频情感分类》。Multimedia Tools and Applications，80(2):2887–2905，2021。7，9[27]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《通过自然语言监督学习可迁移视觉模型》。载于《国际机器学习会议论文集》，第 8748–8763 页。PMLR，2021。8[28] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。《通过大规模弱监督实现稳健的语音识别》，2022。4，5[29] Lala Shakti Swarup Ray、Bo Zhou、Sungho Suh 和 Paul Lukowicz。《基于传感器的开放词汇活动识别：文本嵌入反演的初步发现》。载于《2025 年 IEEE 普适计算与通信国际会议研讨会及其他附属活动（PerCom研讨会）论文集》，第 685–688 页。IEEE，2025。8[30] Karen L Schmidt 和 Jeffrey F Cohn。《人类面部表情作为适应：面部表情研究中的进化问题》。American Journal of Physical Anthropology：美国体质人类学家协会官方刊物，116(S33):3–24，2001。1[31] Tim J Schoonbeek、Tim Houben、Hans Onvlee、Fons Van der Sommen 等。《IndustReal：用于在类工业场景的第一人称视频中处理执行错误的程序步骤识别数据集》。载于《IEEE/CVF冬季计算机视觉应用会议论文集》，第 4365–4374 页，2024。2[32] Fadime Sener、Dibyadip Chatterjee、Daniel Shelepov、Kun He、Dipika Singhania、Robert Wang 和 Angela Yao。《Assembly101：用于理解程序性活动的大规模多视角视频数据集》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 21096–21106 页，2022。2，1[33] Satya P Singh、Madan Kumar Sharma、Aim´e Lay-Ekuakille、Deepak Gangwar 和 Sukrit Gupta。《结合自注意力的深度卷积长短期记忆网络：利用可穿戴传感器进行人类活动解码》。IEEE Sensors Journal，21(6):8575–8582，2020。7，9[34] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。《Qwen2-VL：增强视觉语言模型在任意分辨率下对世界的感知》。arXiv 预印本 arXiv:2409.12191，2024。8[35] Xiao Wang、Zongzhen Wu、Bo Jiang、Zhimin Bao、Lin Zhu、Guoqi Li、Yao wei Wang 和 Yonghong Tian。《HardVS：Re-

利用动态视觉传感器进行人类活动识别。在 *AAAI* 人工智能会议论文集，第 5615–5623 页，2024 年。1[36] Yuanyuan Xu、Wan Yan、Haixin Sun、Genke Yang 和 Jil- iang Luo。Centerface：将人脸视为点进行人脸检测与对齐。在*arXiv:1911.03599*，2019 年。5[37] Naoya Yoshimura、Jaime Morales、Takuya Maekawa 和 Takahiro Hara。OpenPack：用于物联网物流环境中包装作业识别的大规模数据集。在 *2024 IEEE* 国际普适计算与通信会议（*PerCom*），第 90–97 页。IEEE，2024 年。1,2,7[38] Hao Zheng、Regina Lee 和 Yuqian Lu。HA-VID：用于全面理解装配知识的人体装配视频数据集。神经信息处理系统进展，36:67069–67081，2023 年。1, 2

OpenMarcie：工业环境多模态动作识别数据集

补充材料

本补充文档包含有关 OpenMarcie 的更多信息。第 6节详细比较了先前数据集，阐述了传感覆盖范围、工业真实性和标注设计方面的差异。第 7 节分别介绍了每位参与者的元数据信息。第 8 节阐述了 OpenMarcie 数据集的伦理考量和社会影响，重点关注用户隐私和匿名化。第 9 节介绍了大语言模型如何对人工撰写的描述进行语义转换，反之亦然；该模型充当结构化翻译器，而非主要标注工具。第 10 节展示了两个实验场景中的物体分布，并通过示例场景展示物体分割和人体骨架姿态。第 11节报告了自行车场景中的工具操作统计数据，并重点介绍了工具接触事件所涉及的多模态线索。第 12 节详细介绍了所提出基准测试的实现，并展示了所用的评估指标。第 13 节报告了额外的纯音频实验，这些实验旨在探究该模态在不同条件下的局限性，并分析音频在工具接触检测中的作用。具体而言，我们（1）使用匿名化处理前的（原始）音频流评估完整流水线，以单独考察匿名化处理的影响；（2）用非机器学习的梅尔频谱图分类器取代 EnCodec 嵌入，以区分嵌入方法与内容本身的影响。第 14节说明了所选传感器模态及其互补物理信号背后的依据。第 15 节报告了现代多模态架构的结果，包括 Perceiver IO 和 Any-MAL，并展示了相较于后期融合基线的性能提升。第 16节概述了若干超越本文所述基准测试的研究方向。

6. 与先前数据集的详细比较、范围与泛化

为提供更清晰的视角，我们对每个引用的数据集进行叙述性比较，重点说明其设计选择在工业真实性、感知广度以及对高级研究任务的适用性方面，与 OpenMarcie 有哪些趋同或差异。

• InHARD [11], 专为人机协作开发，整合了三个外向视角 RGB 视图与可穿戴惯性动作捕捉设备，使其成为少数纳入可穿戴设备感知的工业数据集之一。然而，其范围仅限于固定外向视角设置下的单动作识别。相比之下，OpenMarcie 更进一步，将可穿戴设备与同步的第一人称和第三人称视频相结合，并提供多动作标签，从而捕捉重叠活动，更真实地反映现实世界中的工业工作流程。

• LARa [24] 使用惯性测量单元、光学动作捕捉和单台 RGB 摄像机采集物流活动，为了解仓库环境中的工人差异性提供了宝贵见解。然而，其研究范围主要局限于第三人称视角，也缺乏 OpenMarcie 所具备的多模态深度。由于没有第一人称视频流或同步的多动作标签，LARa 最适用于受控的人类活动识别场景。OpenMarcie 通过整合第一人称视频、可穿戴传感和多路线任务设计，拓展了这一领域，从而为协作和多任务情境提供更丰富的监督信息。

• OpenPack [37]是一个大规模物流数据集，包含 53.8 小时的录制数据，整合了惯性测量单元、二维关键点、深度和物联网数据。其主要优势在于将可穿戴传感与流程级日志记录相结合，但视角仍完全局限于第三人称。OpenMarcie 通过提供与第三人称多视角画面对齐的第一人称多模态视频，并引入明确的多动作标注，弥补了这些不足。这种设计不仅使模型能够对活动进行分类，还能推理活动的同时性和任务意图。

• Assembly101 [32] 提供多视角第一人称和第三人称视频及密集程序标注，为基于视觉的活动理解建立了有力的基准测试。然而，它未纳入可穿戴传感，也不包含重叠活动标签。OpenMarcie 在拓展这一方向的同时保留了第一人称与第三人称视角覆盖，并新增可穿戴惯性测量单元、音频、热成像数据和光谱仪信号，还明确对并发动作进行建模。这些新增内容使 OpenMarcie 特别适合推动跨模态学习。

• IKEA-ASM [7] 是一个家具组装数据集，提供 RGB-D 多视角录像，并标注了原子动作和被操作物体。其主要优势在于包含来自多个第三人称 Kinect 传感器的深度和姿态数据，但缺少第一人称视角和可穿戴传感。OpenMarcie 通过结合第一人称视频、可穿戴设备及其他非视觉传感器，对这一资源形成补充，从而支持超越单纯 RGB-D 的跨模态迁移。

• HA4M [9] 专注于齿轮传动组装，并提供使用 Azure Kinect 采集的六种模态，包括 RGB、深度、红外和骨架数据。虽然它强调

视觉信息丰富的多模态，但它仅限于第三人称视角，不包含可穿戴传感。Open- Marcie 通过整合第一人称和第三人称视角摄像机、可穿戴设备及多动作标签，拓展了这一领域，从而支持工业工作流程中更贴近现实的多任务处理场景。

• HA-ViD [38] 提供细粒度的多视角装配录像，并配有涵盖主体、动词、物体、工具及协作线索的详细标注。它提供丰富的语义监督，并支持替代任务路径。然而，它不包含可穿戴传感或第一人称视角。Open Marcie 通过引入第一人称视频、可穿戴数据流以及显式的并发多动作标签来补充 HA-ViD，从而将适用性扩展至多任务处理与多模态对齐。

• IndustReal [31] 专为工序步骤识别（PSR）而设计，利用第一人称视频捕捉程序性错误和灵活的子目标。其独特贡献在于聚焦误差建模，但它仍局限于仅自我视角的录像，不具备可穿戴传感或多视角第三人称视角支持。OpenMarcie 目前虽在人类活动识别、图像描述生成和跨模态对齐上进行基准测试，但整合了第一人称视频与第三人称视角视频、可穿戴设备以及多动作标签，为 PSR 和误差建模提供了互补基础——这两者均已纳入我们的分阶段路线图。

• Ego-Exo4D [17] 是一个大规模数据集，包含超过 1,200 小时的熟练活动数据，采集内容涵盖同步的第一人称和第三人称视频、音频、惯性测量单元、视线及语言流。其规模和模态广度无可比拟，但其中只有一小部分记录来自工业场景。OpenMarcie 采取互补性方法，专门聚焦于工业工作流程，并通过多动作标注和专用可穿戴设备对其进行增强。这样，OpenMarcie 成为了 Ego-Exo4D 面向特定领域的对应数据集，在工厂场景中优先追求深度而非规模。

OpenMarcie 是唯一一个同时提供可穿戴设备、第一人称视角与第三人称视角的多视角记录、明确的并发多动作标签，以及完整工业场景覆盖的数据集（见表 1）。先前数据集各有侧重，例如丰富的语义标注（HA-ViD）、大规模数据（Ego-Exo4D）、可穿戴设备集成（InHARD、LARa、OpenPack）或多视角视频（Assembly101、IKEA-ASM），但没有一个能在工业环境中兼具所有这些要素。凭借这一独特定位，OpenMarcie 尤其适用于我们路线图中提出的高级基准测试，包括程序化规划、技能评估、意图预测、细粒度分割、姿态推理、跨模态迁移和跨模态生成。

6.1. 范围与工业定位

OpenMarcie 专注于以装配为核心的工业工作流程，而非试图全面涵盖所有工业活动。该数据集围绕两种互补模式构建：（i）依靠经验的临时装配与拆卸（自行车），记录探索、纠错行为和目标导向的问题解决过程；（ii）遵循指令的程序化装配（3D打印机），体现标准化生产和入职培训场景。

虽然我们并不声称涵盖大规模包装、重型机械检查或基于传送带的物流等活动，但该程序性场景自然纳入了装配之外与工业相关的行为，包括拆包、零件整理、阅读手册、工作空间准备、物体运输，以及对部分完成任务的顺序式协作延续。这些行为反映了受控环境中的结构化生产线动态。

这一设计选择优先考虑可控的多样性和丰富的多模态信息，而非最大限度地拓宽任务范围，从而能够系统地研究装配工作流程中的模态互补性和程序性推理。

6.2. 生态效度与受控测试台环境

数据采集是在受控的测试台环境中进行的，而非运营中的工厂。因此，重型机械噪声、大规模环境杂乱、严格的安全约束以及计划外的工作流程中断等因素，在数据中的体现不足。我们承认这是一个局限。

然而，受控环境也带来了若干方法上的优势：同步多模态采集、减少混杂变量、稳定的相机校准以及较高的标注保真度。这种设置能够实现不同模态之间的精确时间对齐，并系统地研究跨模态鲁棒性；若在不受控的工厂环境中进行，这些工作将困难得多。

因此，OpenMarcie 应被视为一个受控但以工业场景为灵感的装配场景多模态感知基准测试，而非对真实工厂车间的直接复刻。

6.3. 泛化与领域偏移的考量

关于泛化，在 Open-Marcie 上训练的模型预计最直接地迁移到面向装配的任务，这些任务涉及精细运动操作、程序性推理，以及在中等环境变化下的多模态感知。迁移到截然不同的工业领域——例如高噪声加工环境、高速包装流水线，或大规模质量检验流水线——

对于高速包装线或大规模质量检测流水线等场景，无法保证适用性。

未来的评估应明确考察领域偏移条件下的鲁棒性，包括模拟机械噪声注入、环境杂乱增强，以及在运营中的工厂环境中开展跨场地验证。此类实验将为现实世界中的可迁移性提供定量评估。

在此背景下，OpenMarcie可作为研究多模态鲁棒性、模态互补性，以及以装配为中心的工业工作流程中程序生成式活动建模的结构化基础，同时也认识到其生态效度的边界。

7. 用户元数据

基于图 8、表 6和表 7，两个场景中的参与者数据——自行车组装（场景（a））和3D打印机组装（场景（b））——在人口统计特征和职业背景方面反映出一个多样化且较为均衡的群体。场景（a）包括12名参与者，而场景（b）有24名，为分析提供了更广泛的基础。男性在两组中均占多数，不过场景（b）表现出更高的性别多样性，女性参与者人数更多。

如汇总后的可视化结果所示，参与者大多为右利手，仅有四人自认为是左利手，这表明该数据集中右利手占据绝对多数。参与者年龄从二十岁出头到三十多岁不等，大多数拥有工程学学位。其他学科背景还包括计算机科学、生物学、物理学和管理学，体现了该数据集的多学科相关性。

从地理分布来看，该数据集涵盖来自多个大洲的参与者，包括南美洲、亚洲、欧洲、非洲和北美洲，体现了广泛的国际代表性。场景（b）的人口统计特征尤其丰富，涵盖了 15 种以上不同的国籍来源。参与者根据自我报告，以 1 到 3 的等级评估经验水平；不同参与者的经验各异，大多数表示对该任务领域具有初学者到中级程度的熟悉度。

总体而言，这些元数据展现了参与者群体的包容性和多样性，有助于支持该数据集在具身人工智能、人机交互和基于视觉的行为分析领域中开发可泛化模型的用途

sis.

除人口统计信息外，我们还报告了场景（b）的上半身人体测量数据，汇总见表8。

这些测量值包括肩宽（SS）和肩至腕长度（SR-WR 和 SL-WL），单位为厘米，可对参与者的身体比例进行定量描述。

视频中人物的身体尺寸，尤其是其绝对尺度、边界框大小和相对肢段比例，与姿态估计性能密切相关。尽管许多现代算法力求实现尺度不变性，但身体的像素级表示会直接影响关键点定位精度，进而影响三维重建。在我们的设置中，摄像机安装在环境中，用于估计参与者执行组装任务并在房间内自由移动时的姿态；参与者与摄像机的距离不断变化，因此其表观尺度也随之改变。因此，提供特定于参与者的人体测量值，可在以厘米为单位的参考坐标系中实现更精确的度量重建，有助于准确进行多摄像机三角测量，并为评估和改进姿态估计算法提供可靠的真实标注信息。

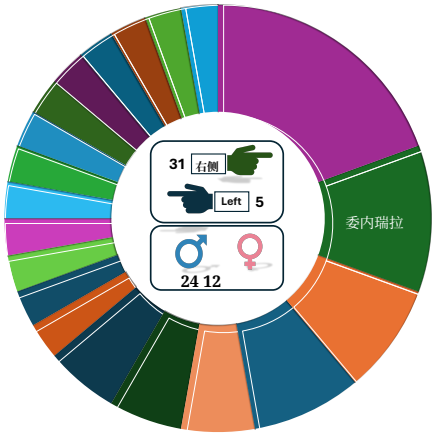


图8。按国籍、优势手和自我认定性别划分的参与者分布。

图9 比较分析了两个不同装配场景中的运动范围（左侧）和运动多样性（右侧）：场景（a）自行车装配，以及场景（b）3D打印机组装。每项指标均按身体区域细分：全身、左手、右手和下半身。与场景（a）相比，场景（b）的运动范围更大，尤其是全身以及左手和右手。这种动作增多可能是因为在场景（b）中多次出现了围绕桌子、架子和地板进行搜索的情况。场景（a）则包括跪姿、躺在地上以及站起检查自行车的特定部位。因此，与3D打印机场景（场景b）相比，场景（a）的下半身运动范围和动作多样性更大。

表9 列出了用于表示场景（a）：自行车组装中涉及的各种活动和物体的首字母缩略词。活动以单个字母表示

表6. 临时场景（a）中参与者的元数据信息。

ID	Sex	Age	身高	优势手	学术水平	人口统计信息	经验水平（1-3）
P1-B	F	33	160	R	工程师	南美洲（委内瑞拉）	2
P2-B	F	26	176	R	工程师	欧洲（波兰）	3
P3-B	M	29	175	R	计算机科学家	欧洲（德国）	1
P4-B	M	26	175	L	工程师	南亚（印度）	1
P5-B	M	33	189	R	计算机科学家	南美洲（巴西）	1
P6-B	M	36	188	R	工程师	东非（卢旺达）	1
P7-B	M	25	168	R	工程师	南亚（印度）	1
P8-B	M	37	178	R	工程师	东亚（韩国）	2
P9-B	M	27	176	R	工程师	中东（伊朗）	2
P10-B	M	27	174	R	工程师	南亚(印度)	1
P11-B	M	34	193	R	工程师	南美洲 (委内瑞拉)	1
P12-B	M	30	160	R	工程师	东南亚（中国）	1

表7. 程序性场景（b）的参与者元数据信息。

ID	性别	Age	身高	优势手	学术水平	人口统计信息	经验水平（1-3）
P1-D	M	37	178	R	工程师	东亚（韩国）	2
P2-D	F	25	159	R	工程师	东南亚及太平洋地区（尼泊尔）	2
P3-D	M	34	193	R	工程师	南美洲（委内瑞拉）	1
P4-D	F	26	165	R	工程师	欧洲（意大利）	1
P5-D	F	33	160	R	工程师	南美洲（委内瑞拉）	2
P6-D	F	25	164	R	工程师	南亚（印度）	1
P7-D	M	24	175	R	工程师	中东（伊拉克）	3
P8-D	M	25	168	R	工程师	南亚（印度）	1
P9-D	F	25	162	R	物理学家	南美洲（哥伦比亚）	2
P10-D	F	27	155	R	工程师	东南亚（印度尼西亚）	2
P11-D	F	24	160	R	工程师	南亚（印度）	1
P12-D	M	36	188	R	工程师	东非（卢旺达）	1
P13-D	M	24	168	R	生物学家	北美洲（美国）	2
P14-D	M	22	187	R	管理	中东（叙利亚）	2
P15-D	F	24	150	R	工程师	高加索亚洲（阿塞拜疆）	1
P16-D	M	25	165	R	工程师	南非（津巴布韦）	1
P17-D	M	22	167	R	工程师	东南亚及太平洋地区（尼泊尔）	2
P18-D	M	29	175	L	计算机科学家	欧洲（德国）	1
P19-D	F	23	161	R	生物学家	东亚（日本）	1
P20-D	M	26	175	L	工程师	南亚（印度）	1
P21-D	M	27	182	R	计算机科学家	东亚（韩国）	2
P22-D	M	24	183	L	工程师	东南亚及太平洋地区（尼泊尔）	2
P23-D	F	24	156	R	微生物学家	南美洲（哥伦比亚）	1
P24-D	M	25	178	L	计算机科学家	中东（土耳其）	1

缩略词，例如 W 表示行走，M 表示移动（在下一步动作之前操作物体），U 表示拧紧/拧松，C 表示骑自行车。其他肢体动作包括 S 表示坐下，P 表示给轮胎打气，I 表示用手或眼睛检查，H 表示锤击，K 表示跪下，A 表示站起，L 表示躺下，T 表示剪切。此外，表中还列出了与物体相关的缩略词，包括 hx 表示内六角扳手，wr 表示扳手，sd 表示螺丝刀，hm 表示锤子，sc 表示剪刀，pl 表示钳子，pu 表示泵，bh 表示徒手。该表有助于阐明用于描述自行车组装过程中详细步骤和所用工具的简写。

8. 伦理考量与社会影响

OpenMarcie数据集的开发高度重视符合伦理的研究实践和社会责任。所有参与者均按照《赫尔辛基宣言》的规定提供了知情同意；该研究由德国人工智能研究中心的伦理委员会根据 HRW-35/24 和 SMD-30/24 方案进行审查并获批。参与者已获知所收集数据的类型，包括第一人称视频、可穿戴传感器数据和外部音频叙述，并保留随时退出的权利。参与者自愿参与后获得了价值 15 欧元的亚马逊代金券

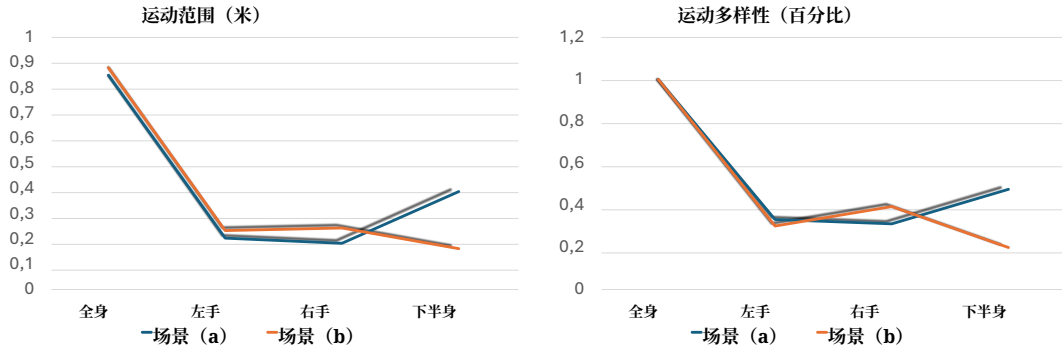


图9。身体动作范围 a 以及场景（a）：自行车组装和场景（b）：3D打印机装配中的身体动作多样性 y.

表8. 上半身人体测量数据：肩宽（SS）、右肩至右腕长度（SR-WR）以及左肩至左腕长度（SL-WL），单位为厘米。表中列出每位参与者的数值，以及均值 ±标准差。

ID	SS（厘米）	SR-WR（厘米）	SL-WL（厘米）
P1-D	51	55	55
P2-D	34	49	48
P3-D	51	61	61
P4-D	36	57	57
P5-D	36	54	54
P6-D	38	53	53
P7-D	46	59	59
P8-D	44	55	54
P9-D	40	55	54
P10-D	36	50	50
P11-D	38	53	53
P12-D	47	65	64
P13-D	44	55	55
P14-D	50	61	61
P15-D	38	46	46
P16-D	45	54	54
P17-D	41	56	57
P18-D	43	56	56
P19-D	32	52	52
P20-D	42	56	56
P21-D	47	59	59
P22-D	47	57	58
P23-D	38	49	49
P24-D	42	57	57
均值 ± 标准差	41.33 ± 5.82	55.04 ± 4.42	54.96 ± 4.43

参与场景（b）。参与场景（a）不提供任何报酬。

为降低隐私风险，第一人称视角录制经过刻意取景，以捕捉任务执行过程，同时尽量减少参与者身份的暴露。口头叙述由外部观察者提供，以避免采集参与者的私人言语。随后使用 faster-whisper（v1.1.1）和 ctranslate2（v4.4.0）对这些叙述进行转录，并采用“large-v3”模型配置 [28], , 基准测试任务中仅使用所得文本描述，并将其发布于 OpenMarcie。

面部特征与生物识别标识符被排除在所有标注和分析流程之外。对视频数据应用了两阶段匿名化流程。首先，使用开源 deface Python 软件包自动检测并模糊参与者面部（阈值设为 0.7） [25, 36]。其次，对所有视频进行人工复核，并使用 DaVinci Resolve 将任何残留的可见面部特征完全遮盖 [13]。对于音频，仅发布伴奏部分。使用 Audacity 中的 OpenVINO 音乐分离插件有意从音轨中去除人声 [19], , 该插件可将录制内容分离为人声与伴奏分轨。OpenMarcie 仅包含伴奏音轨。

OpenMarcie 不仅旨在推动人类活动识别研究，也致力于促进机器学习的透明度和公平性。该数据集包含参与者人口统计特征、学术背景、优势手以及自评技能水平等全面的元数据，使研究人员能够开展亚组分析并审查潜在偏差。这有助于在工业和具身人工智能应用中开发公平的多模态系统。

从更广泛的社会层面来看，OpenMarcie 有望惠及人机协作、工作场所安全、人体工学评估和自适应培训系统等应用。其贴近现实、以目标为导向的场景非常适合推动能够解读复杂环境中人类动作的模型发展。然而，如果使用不当，此类系统也可能带来风险，例如在未经同意的情况下进行过度监控或绩效监测。

我们鼓励使用 OpenMarcie 的研究人员明确评估公平性，记录不同子群体的表现，并考虑在以人为本的场景中部署人类动作识别系统所带来的后续影响。OpenMarcie 旨在通过提供丰富且符合伦理原则的现实世界多模态学习测试平台，支持人工智能的负责任发展。

表9. 场景（a）：自行车组装中的活动和物体首字母缩略词。

缩略词	含义
W	行走
M	移动：操作物体，直至下一个动作
U	拧紧/拧松
C	骑自行车
S	坐下
P	泵：给自行车轮胎打气
I	检查：用手/眼睛检查物体
H	锤击
K	跪下
A	站起
L	躺下
T	切割
hx	内六角扳手
wr	扳手
sd	螺丝刀
hm	锤子
sc	剪刀
pl	钳子
pu	Pump
bh	徒手

9. 基于大语言模型的标注翻译和真实标注策略

9.1. 大语言模型生成标签验证

我们的流水线使用 GPT-4o 将人工编写的软活动描述转换为标准格式——离散的活动类别（硬标签）或连续表征（软标签）。重要的是，大语言模型并非用于直接根据视觉输入生成标注，而是对现有的人工编写标注进行语义转换。因此，它充当的是结构化翻译器，而非主要的标注代理。

重要的是，这一转换过程并非完全自动化。整个过程中都融入了人工监督，尤其是在将软标签句子映射到离散类别时。标注人员会优化大语言模型提示，检查输出中的模糊情况，并解决语义不一致问题。这种人在回路方法有助于确保用于模型训练的最终标签准确且一致。

为验证大语言模型生成的标注质量，我们在以下两种场景下进行双向一致性分析：

- **临时场景（硬标签 →描述 → 硬标签）：** 人工标注的离散类别由大语言模型转换为自然语言描述，再由大语言模型反向转换为离散标签。
- **程序生成场景（描述 → 硬标签 → 描述）：** 人工撰写的软标签句子由大语言模型转换为离散类别，再重新生成描述。

我们使用宏平均 F1 分数和 METEOR 来衡量原始标签/说明与恢复后的标签/说明之间的一致性[2]。在两种设置下，我们都观察到分类与回归输出之间具有很强的一致性——场景（a）的**宏平均 F1 分数为 0.715**，场景（b）的**METEOR 分数为 0.531**，这表明大语言模型生成的标签保留了语义一致性和结构保真度。这种间接验证证实了大语言模型在部分监督标注流程中作为可靠语义中介的效用。

9.2. 临时场景（a）：自行车拆卸/装配

例如，给定：

动词：移动物体和行走 工具：徒手 被操作物体：内六角扳手 备注：“将物体从桌子移向自行车”

我们发出以下提示：

% 系统消息，用于定义助手角色和风格 系统：你是一位专业的活动描述助手。始终使用现在进行时、第三人称，输出一个以大写字母开头并以句号结尾的陈述句。

% 包含标注载荷和示例的用户消息

用户：
将以下结构化标注转换成一个清晰的句子。

标注：
{ 动词：移动物体和行走 { 工具：徒手 { 被操作物体：内六角扳手 { 备注：将物体从桌子移向自行车

现在，请将上述标注转换成句子。

GPT-4o 随后输出：

\他正徒手移动内六角扳手并行走。"

此生成句子用作下游模型训练的软标签目标。

9.3. 程序生成场景（b）：3D打印机装配/拆卸

在第一阶段，我们使用推理模型 DeepSeeker1[18] 对所有软标签句子预测候选活动类别。我们迭代提取预测类别的唯一集合，并反复运行流水线，直到收敛（即不再出现新类别）。接下来，我们人工核验并合并语义相似的类别，生成最终的离散集合：拾取、移动、抬起、行走、坐、调整、站立、阅读、投掷、饮用、弯腰。

例如，给定软标签：

\人物弯腰从地板上拾取某物。
"

我们构造一个提示词，引导 DeepSeeker1 针对所描述的动作进行开放式推理，例如：

系统：
你是一名推理助手，负责从描述人类行为的自然语言中提取与活动相关的动词或动作短语。用户：从以下句子中提取关键活动标签：\这个人弯腰从地板上拾取某样东西。"

DeepSeeker1 随后可能返回：

[“拾取”、“弯腰”、“抓取”]

每个返回的标签都被视为一个候选活动类别。我们会逐一将其与当前已知类别的工作集进行核对。如果某个标签尚未包含在集合中，就将其添加进去。我们会针对整个软标签句子数据集反复执行这一迭代过程。每轮结束后，我们会对任何新发现或含义模糊的短语重新运行 DeepSeeker1，以发现可能遗漏的类别。该过程将持续进行，直到收敛，也就是说，完成一轮完整迭代后，不再添加新的唯一类别。

在我们的示例中，如果“拾取”和“抓取”已在工作集中，但“弯腰”不在，我们就会将集合更新为：

现有类别集合： [“拾取”、“抓取”，.....]更新后的类别集合： [“拾取”、“抓取”、“弯腰”，...
...]

收敛后，我们会手动核验并合并语义相似或冗余的标签（例如合并“弯腰”和“弯腰”），以最终确定一个整洁的活动类别离散集合：

最终类别集： [“拾取”、“移动”、“抬起”、“行走”、“坐”、“调整”、“站立”、“阅读”、“投掷”、“饮用”、“弯腰”]

在第二阶段，我们使用 GPT-4o [1] 通过提示工程，将每个软标签句子转换为上述类别之一，并采用“粘滞”逻辑，使预测类别持续有效，直到在之后的某个时间戳检测到新的类别。

例如，给定软标签：“人物正坐在椅子上，位于桌子前方。”
我们给出以下提示：

系统：
你是一名活动分类助手。候选类别：拾取、移动、抬起、行走、坐、调整、站立、阅读、投掷、饮用、弯腰、其他。粘滞逻辑：除非预测出新标签，否则保留上一个标签。用户：将以下句子分类为某个候选类别：\人物坐在椅子上，位于桌子前方。"

表 10. 使用 VLM自动标签与人工标注进行训练时的开放词汇余弦相似度。VLM标签的得分仅略高于随机水平，显著低于手册基线。

模态	场景（a）		场景（b）	
	余弦相似度（↑）			
	无空类	Null	无空类	Null
I	0.137±0.041	0.129±0.038	0.168±0.012	0.164±0.011
A	0.098±0.035	0.091±0.033	0.103±0.009	0.099±0.010
V	0.152±0.029	0.143±0.027	0.181±0.008	0.177±0.009
I + A	0.141±0.039	0.133±0.036	0.172±0.010	0.168±0.011
A +V	0.148±0.033	0.139±0.030	0.179±0.009	0.174±0.009
I + V	0.163±0.028	0.154±0.026	0.194±0.007	0.189±0.008
I + A +V	0.159±0.031	0.150±0.029	0.190±0.008	0.185±0.009

GPT-4o 随后输出：“10（坐）”

其中整数“10”指的是类别“坐”。其中整数“10”指的是类别“坐”。这种两阶段方法为下游模型训练生成了最终的硬标签。

真实标注获取策略。 上述真实标注依赖同步多视角拍摄，而非专用动作捕捉系统。尽管高精度光学动作捕捉原则上能够提供精确的运动学测量，我们还是有意避免采用基于标记点的设置。这类系统不适用于长时间跨度的工具使用，可能会限制自然操作，且在涉及频繁手物交互的近距离装配过程中尤其容易受到遮挡影响。这些限制可能损害以装配为中心的工作流程中的生态效度。为此，我们采用 ZED 立体相机进行同步多视角 RGB-D 采集，在尽量减少干扰的同时实现强大的三维场景感知，并具备良好的可扩展性。人工意图感知标注在信息量最大的外中心视角上进行，并在各模态间进行时间对齐，从而确保真实标注在语义上一致，同时保留自然的交互行为。

基于视觉语言模型的自动标注 为评估视觉语言模型（VLM）能否取代或减少人工标注工作，我们通过提示最先进的视觉语言模型 Qwen2-VL，生成自动软标签 [34]，并使用与人工标注相同的时间边界为每个自中心视频片段生成描述。表 10报告了使用这些由视觉语言模型生成的标签（而非人工撰写的软标签）训练回归流程时取得的开放词汇余弦相似度。

在所有模态组合和两种场景中，VLM自动标签的余弦相似度分数范围为 0.09–0.19，仅略高于随机嵌入基线（≈0.0）。相比之下，使用人工编写的软标签训练的同一回归模型，在场景（a）中达到 0.36–0.56，在场景（b）中达到 0.32–0.66，差距达 3–4×。定性检查发现，VLM生成的描述往往聚焦于泛化的场景背景（“一位人物站在工作间里”），而不是人工标注者记录的细粒度工具与物体交互（“他正用内六角扳手拧紧刹车”）。在较短的动作片段中，这种差异尤为明显，因为 VLM 缺乏足够的时间上下文，难以区分视觉上相似的活动。鉴于这些发现，我们未将 VLM自动标签纳入主要基准测试。

表 11. 场景（a）（自行车组装）的工具词汇。百分比根据所有用户合计的 1,769 个带工具标注的动作实例计算。

Code	Tool	使用率 %
bh	徒手	72.5%
hx	内六角扳手	16.2%
sd	螺丝刀	3.3%
wr	扳手	2.9%
hk	Hook	2.2%
pu	Pump	1.0%
pl	钳子	1.0%
hm	锤子	0.6%
sc	剪刀	0.2%

10. 目标跟踪

OpenMarcie提供了多视角的物体分割与跟踪数据，包括用于3D组装实验（场景（b））的一台外视相机和两个第一人称视角位置（安装在头部和胸部的可穿戴相机）。图10 上 展示了从外部视角检测到的场景（a）自行车组装和场景（b）3D打印机组装中的物体分布。在这两种情况下，“人物”都是最常被观察到的类别，反映了参与者在执行任务期间始终在场。笔记本电脑、手机和包等常见物体在两种场景中均有出现，这可能体现了工作场景中常见的配件。

特定场景中的物品凸显了背景差异：例如，自行车仅出现在自行车组装场景中，而箱子——可能代表包装或组件——则仅出现在3D打印机组装场景中。运动球和狗等关联较弱或偶然出现的物体检测频率较低。图10下 展示了两个场景中的示例帧，其中包含物体掩码和人体骨架姿态。

总体而言，该图突显了物体存在情况随上下文而变化的特点，为计算机视觉和机器人学应用提供了有价值的见解。

11. 工具操作

桌子 11 列出了场景（a）动词–工具–物体标注方案中使用的九种工具词汇。徒手

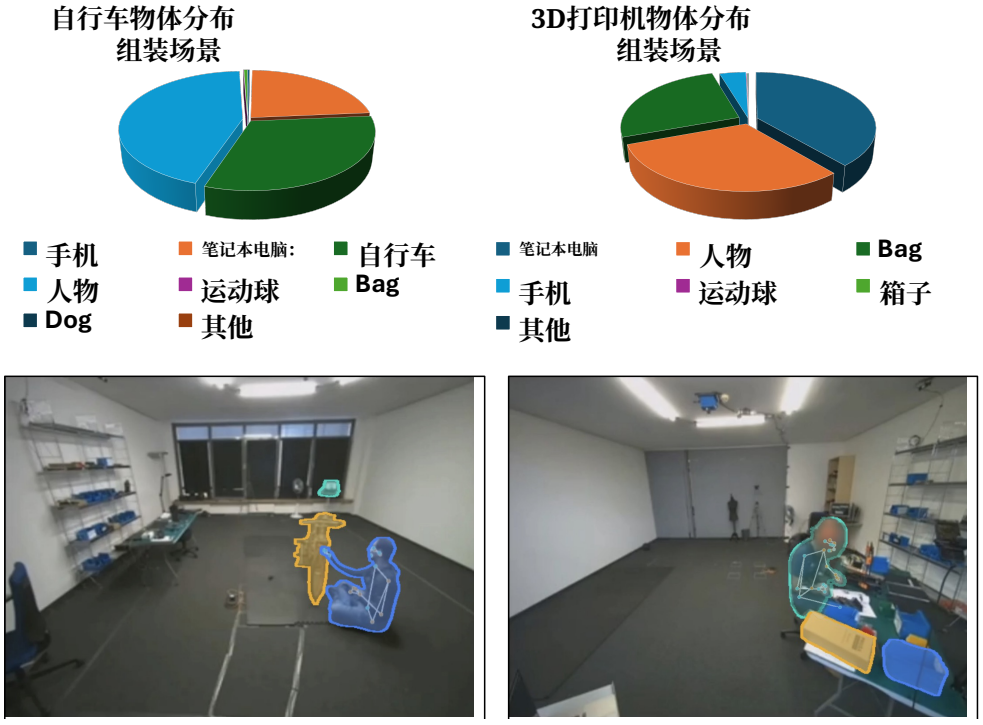


图10。 物体分布 自行车和3D打印机场景中的内容，来自第三人称视角，并配有场景可视化 s
展示物体分割 （掩码）以及这两种环境中的人体姿态估计。

操作占主导（72.5%），反映出自行车装配过程中定位与检查步骤的普遍性。内六角扳手（内六角扳手）占工具实例的16.2%，是最常见的专用工具，这与其在拧紧自行车螺栓时的普遍使用相一致。其余工具——螺丝刀、扳手、撬胎棒、泵、钳子、锤子和剪刀——合计覆盖11.3%的实例。如表18所示，声学模态对于检测这些工具接触事件尤为重要：将音频加入V+I后，宏平均F1在场景（a）中从0.905提升至0.923，在场景（b）中从0.898提升至0.914，因为锤子、扳手和内六角扳手等工具产生的瞬态冲击与摩擦声具有判别性特征，可补充惯性与视觉通道。

12. 基准架构

如图 11所示，我们将数据分割为同步的 1 秒窗口，并评估不同模态组-合在人类活动识别中的表现。对于每种模态，输入数据均独立编码：3 个视频帧$\{x_v^t\}_{t=1}^3$ 通过视觉变换器 [14] $\mathcal{E}_v$进行处理，惯性测量单元信号 $x_i \in \mathbb{R}^{100\times 6}$ 通过 DeepConvLSTM [33]编码器 $\mathcal{E}_i$进行处理，1 秒音频 $x_a \in \mathbb{R}^{16000}$ 通过 En-Codec [12] $\mathcal{E}_a$ 进行处理，生成嵌入 $z_v = \mathcal{E}_v(\{x_v^t\})$。$z_i = \mathcal{E}_i(x_i)$、 $z_a = \mathcal{E}_a(x_a)$，分别如此。对于单模态模型，分类头 $\mathcal{C}$ 将每个 z_m 映射为逻辑值 $\hat{}y = \mathcal{C}(z_m)$，其中 $m \in \{v, i, a\}$。对于多模态组合（例如，视频+惯性测量单元、视频+音频，或三者全部），嵌入通过后期融合 Transformer [26] $\mathcal{F}$进行融合：$z = \mathcal{F}(z_{m_1}, z_{m_2}, \ldots)$，随后得到 $\hat{}y = \mathcal{C}(z)$。所有模型均使用交叉熵损失进行训练：

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{1}$$

其中， $C = 12$ 是活动类别的数量， y 是独热真实标签。

我们将开放词汇描述形式化为句子嵌入回归任务，模型从多模态感官输入中预测语言表征。每种模态均独立处理：3个采样视频帧 $\{x_v^t\}_{t=1}^3$ 经视觉变换器$\mathcal{E}_v$处理，1秒音频 $x_a \in \mathbb{R}^{16000}$ 经EnCodec $\mathcal{E}_a$处理，惯性测量单元信号 $x_i \in \mathbb{R}^{100\times 6}$ 经DeepConvLSTM $\mathcal{E}_i$处理，从而生成嵌入 $z_v = \mathcal{E}_v(\{x_v^t\})$、 $z_a = \mathcal{E}_a(x_a)$和$z_i = \mathcal{E}_i(x_i)$。对于单模态或多模态组合，嵌入经由 Transformer $\mathcal{F}$融合，得到共享表征 $z = \mathcal{F}(z_{m_1}, z_{m_2}, \ldots)$。回归头 $\mathcal{R}$ 将 z 映射为预测的句子嵌入$\hat{s} = \mathcal{R}(z)$。真实句子嵌入 s由一个

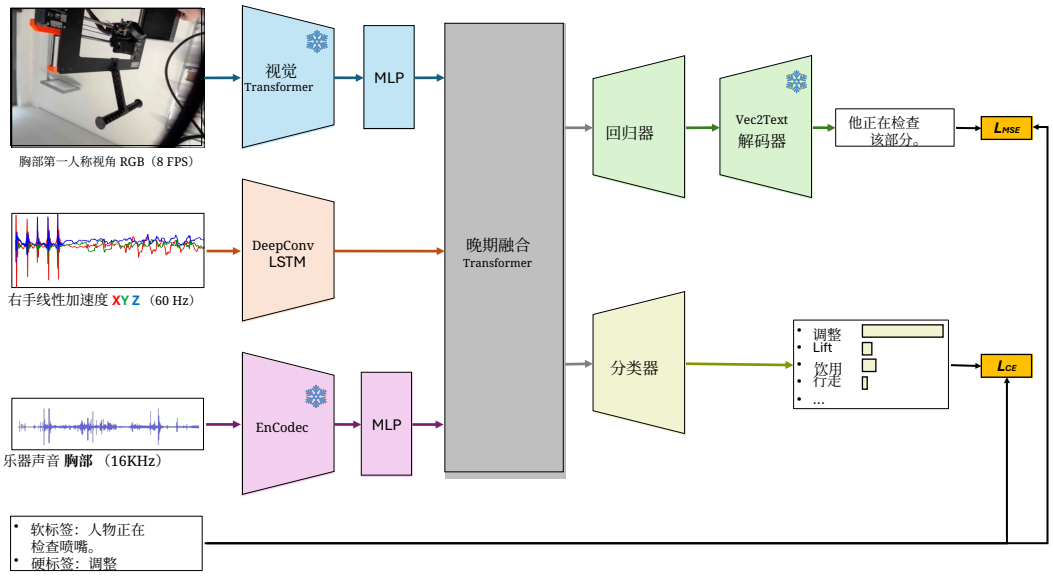


图11. 用于 人类活动识别与开放词汇图像描述基准测试中的分类和回归 s.

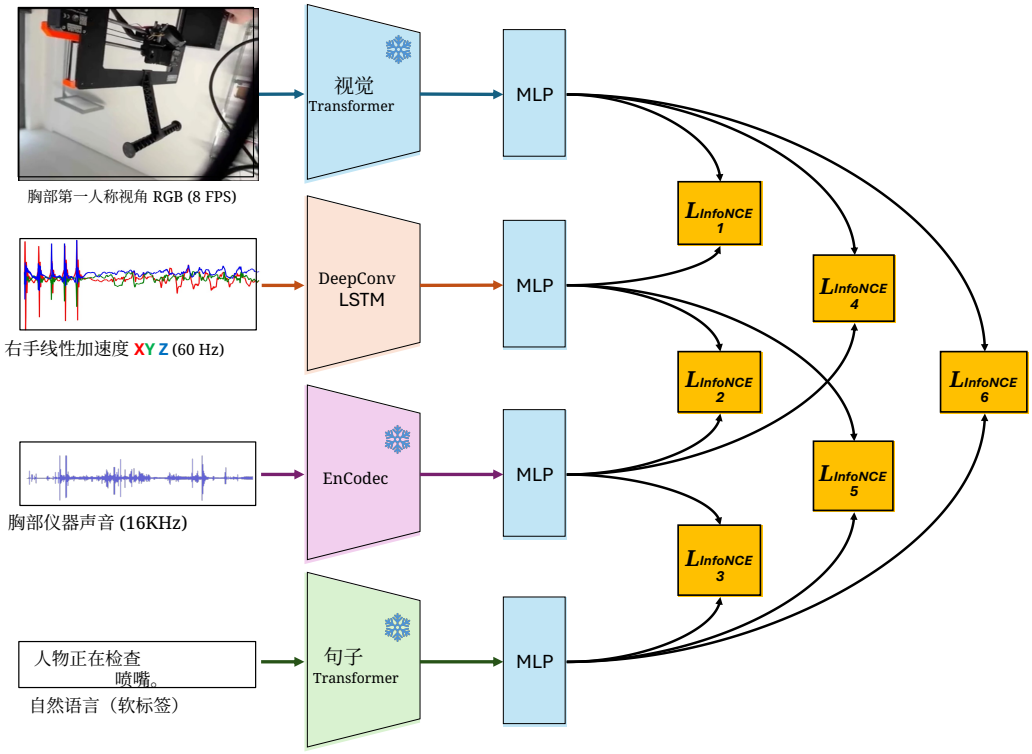


图12. 跨模态对齐基准测试的架构。

预训练语言编码器获得。模型经过训练，以最小化均方误差（MSE）：

$$\mathcal{L}_{\text{MSE}} = \|\hat{s} - s\|_2^2 \tag{2}$$

我们使用 Vec2Text [23] 解码器执行图像描述检索。

这种嵌入反演方法能够在不进行自回归解码的情况下，实现可扩展的低延迟图像描述生成。

为了实现跨模态对齐（见图12），我们采用自监督对比学习方法，使用 InfoNCE 损失将不同模态的嵌入映射到

类似于 ImageBind 的共享表示空间[16]。使用与之前相同的模态专属编码器，我们计算每种模态的嵌入 $z_m = \mathcal{E}_m(x_m)$，包括 $m \in$ {视频、音频、惯性测量单元、文本}。对于每个时间对齐对 $(z_m, z_{m'})$，我们应用投影头，并计算其与批次中其他样本的相似度。InfoNCE 损失如下：

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_m, z_{m'})/\tau)}{\sum_{z^-} \exp(\text{sim}(z_m, z^-)/\tau)} \tag{3}$$

其中 sim() 表示余弦相似度，τ 是温度参数，z^- 是负样本。

12.1. 基准测试指标

我们使用适合各项基准测试结构和目标的任务专用指标：

人类活动识别（宏平均 F1 分数）： 为考虑多类别活动识别中的类别不平衡问题，我们报告所有类别的宏平均 F1：

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{4}$$

无论类别频率如何，这种方法都赋予每个类别相同的权重。

图像描述生成（余弦相似度）： 我们通过计算预测句子嵌入与真实句子嵌入之间的余弦相似度来评估描述质量：

$$\text{sim}(\hat{s}, s) = \frac{\hat{s} \cdot s}{\|\hat{s}\| \cdot \|s\|} \tag{5}$$

这反映了模型在共享嵌入空间中捕捉语义相似性的能力，而不依赖精确的词汇匹配。

跨模态对齐（检索指标）： 我们通过检索正确配对的模态来评估对齐情况，使用 Recall@k 和 Top-1 准确率：

$$\text{Recall@}k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[y_i \in \text{Top-}k\left(\hat{y}_i\right)\right] \tag{6}$$

$$\text{Top-1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\arg\max\left(\hat{y}_i\right) = y_i\right] \tag{7}$$

这些指标量化了共享嵌入空间中的检索质量，表明各模态的对齐程度。

为验证双向标注变换的保真度，我们使用两种互补的文本生成指标比较原始标注和恢复后的标注：用于离散标签的宏平均F1分数，以及用于描述的METEOR。

软标签评估（METEOR [2]）： 我们使用METEOR指标比较用户标注的描述 s 和大语言模型生成的描述 $\hat{s}$：

$$\text{METEOR}(s, \hat{s}) = F_\alpha(s, \hat{s})\left(1 - \text{Pen}(s, \hat{s})\right) \tag{8}$$

12.2. 训练细节

所有模型均采用固定的滑动窗口策略进行训练，以确保各模态具有一致的时间上下文。在数据集构建过程中，任何长度短于所需窗口长度的片段都会被跳过，以保证每个样本都具有完整的时间上下文。

所有模型均使用 128 的批量大小和 1×10^{-4} 的初始学习率。晚期融合（分类和回归）模型以及 JLR 对比模型使用 Adam 优化，且不设置权重衰减。所有模型训练 300 轮，并采用早停。学习率调度采用 ReduceLROnPlateau 策略，耐心值为 5 轮，衰减因子为 0.5。

在架构方面，晚期融合和 JLR 模型采用 128 的隐藏维度，包含两层 Transformer 层、四个注意力头，前馈维度为 256。分类模型使用 BCEWithLogits 目标函数，回归模型使用均方误差（MSE），而 JLR 模型采用 InfoNCE 对比目标函数，包含六个成对项和一个温度参数 $\tau = 0.1$。

在权重初始化方面，所有模型均使用 PyTorch 的 Kaiming 均匀初始化来初始化线性层和卷积层，偏差则初始化为零。长短期记忆网络的遗忘门偏置设为 1.0，以促进稳定的长程梯度流。

该框架支持灵活的模态处理。只有通过配置标志启用时，才会实例化可选模态。在批处理期间，系统会获取可选张量；如果某种模态不存在，则返回 None，从而确保在某些数据流不可用时仍具有鲁棒性。在多层感知机编码器隐藏层中的随机失活率为 0.3，在 Transformer 编码器层中的随机失活率为 0.1。

12.3. 计算资源和模型效率

所有实验均在配备 32GB 显存的 NVIDIA RTX 4090 图形处理器上进行。为评估模型的计算开销，我们报告了每项实验的参数总数、乘加运算次数（MACs）和浮点运算次数（每次推理的浮点运算次数），以及相应的推理速度（见表12、表 13和表14）。

13. 音频扩展实验

鉴于声学模态单独发挥的作用有限，我们补充了实验，以便更好地了解其在不同条件下的表现。具体而言，

桌子 12. 跨模态组合联合潜在表示的计算资源。

组合	参数	MACs	浮点运算次数
文本 +音频	781,568	783,616	2.70×10^9
文本 +视频	2,098,432	2,100,480	7.37×10^9
文本 +惯性测量单元	712,512	22,797,312	2.07×10^{10}
文本 +视频 + 惯性测量单元	2,351,552	24,437,376	2.00×10^{10}
音频 +视频 + 文本	2,420,608	2,423,680	5.91×10^9
音频 +文本 + 惯性测量单元	1,034,688	23,120,512	1.89×10^{10}

表13. 各模态组合分类模型的计算资源。

组合	参数量	MACs	浮点运算次数
仅音频	339,200	340,224	2.72×10^9
仅视频	1,656,064	1,657,088	1.38×10^{10}
仅惯性测量单元	270,144	22,353,920	2.39×10^{10}
音频 +视频	2,092,928	3,307,392	1.89×10^{10}
音频 +惯性测量单元	707,008	24,004,224	2.44×10^{10}
视频 + 惯性测量单元	2,023,872	25,321,088	2.57×10^{10}
音频 +视频 + 惯性测量单元	2,362,432	26,316,032	2.67×10^{10}

表14. 各模态组合下回归模型的计算资源。

组合	参数量	MACs	浮点运算次数
仅音频	746,624	744,000	5.95×10^9
仅视频	1,952,768	1,952,000	1.63×10^{10}
仅惯性测量单元	566,848	22,648,832	2.42×10^{10}
音频 +视频	2,270,080	2,268,992	1.70×10^{10}
音频 +惯性测量单元	884,160	22,965,824	2.33×10^{10}
视频 + 惯性测量单元	2,090,368	24,173,824	2.45×10^{10}
音频 +视频 + 惯性测量单元	2,407,616	24,490,816	2.49×10^{10}

我们分析了隐私保护匿名化（移除语音并以器乐音色替代）以及特征提取方法选择（Encodec 嵌入与梅尔频谱图）对三项任务的影响：人体活动识别、开放词汇图像描述生成和跨模态对齐。这些对比使我们能够量化匿名化会在多大程度上降低语义丰富度、其他表示方式能否恢复有用信号，以及音频在多模态场景中仍能发挥多大作用。以下表格报告了这些扩展实验的结果，进一步阐明了声学数据在隐私保护多模态学习中的挑战与潜力。

在人类活动识别（表 15）、开放词汇图像描述生成（表 16）和跨模态对齐（表 17）中，呈现出一致的趋势：未经匿名化的音频表现优于匿名化版本，其中梅尔频谱图的结果最佳。在活动识别任务中，未经匿名化的梅尔频谱图取得最高宏平均F1（场景（a）中为 0.517/0.493，场景（b）中为 0.466/0.455）。在图像描述生成任务中，同一表示方式取得最高余弦相似度（场景（a）中为 0.381/0.359，场景（b）中为 0.330/0.340）。最后，在跨模态对齐任务中，未经匿名化的梅尔频谱图再次以召回率@1/5 和 Top-1分数领先（场景（a）中为 0.254/0.613/0.360；场景（b）中为 0.238/0.597/0.345）。场景（b）的绝对表现始终较低，反映出其程序生成复杂度更高。匿名化与未匿名化数据流之间的性能差距（各项指标相差 0.01–0.03）凸显了隐私与信息丰富度之间的权衡：匿名化会系统性地去除语义丰富度，降低所有任务中的判别能力。尽管如此，音频仍保留了互补线索，尤其是在文本落地方面；即使经过匿名化，跨模态对齐表现仍稳定提升，便可见一斑。

这种模态的局限性源于四个主要因素：

- 隐私约束：移除言语并以通用器乐音色替代，会剥离语境和语义信息。
- 特征提取选择：尽管 Encodec 嵌入用途广泛，但可能无法捕捉细粒度的任务特定线索，尤其是在匿名化信号中。
- 任务固有的挑战：装配声音较为微弱或间歇出现，因此更难区分。
- 环境因素：在测试台环境中采集的数据缺乏真实的工业声学特征（例如机械噪声、振动），这进一步限制了信息量。

总之，声学数据单独使用时表现不佳，甚至可能给分类任务引入噪声，但在多模态融合和联合表征学习中能发挥积极作用，也为探索隐私保护感知提供了测试平台。当前实验证实了更丰富的声学内容具有实用价值，同时也揭示了匿名化和特征选择带来的局限。

使用音频进行接触检测 表18 报告了工具接触/撞击检测结果。二元接触标签是通过将原始开放词汇软标签中的每条时间分段标注分类为**工具接触**或**工具非接触**类别生成的，采用混合流水线：首先通过基于规则的初筛匹配工具相关关键词和动作模式（例如，**锤击**、**用扳手拧紧**），随后通过基于大语言模型的细化步骤处理模糊实例。值得注意的是，尽管单靠声学特征难以胜任一般活动识别，但它们在检测工具接触事件方面信息量丰富，作为单一模态时得分达到 0.783 和 0.772。此外，在最佳双模态组合 V+I 中加入音频，始终能提升工具接触 F1：场景（a）中从 0.905 提升至 0.923，场景（b）中从 0.898 提升至 0.914。这表明，声学通道能够捕捉锤击、咔嗒声和拧螺丝声等瞬态撞击特征；对于粗粒度活动类别，这些特征与惯性测量单元振动大多存在冗余，但当任务要求检测工具与物体之间的物理接触时，它们便具有区分性

表15. 音频扩展实验中场景（a）：自行车组装和场景（b）：3D打印机组装的人类活动识别结果，包括有空类和无空类时的宏平均F1分数。

模态	场景（a）		场景（b）	
	宏平均 F1（↑）			
	不含空类	带空类	不带空类	带空类
声学匿名化编码器	0.489±0.018	0.469±0.017	0.425±0.004	0.432±0.005
声学非匿名化编码器	0.509±0.002	0.488±0.001	0.460±0.002	0.453±0.003
声学匿名化梅尔频谱	0.492±0.003	0.473±0.002	0.434±0.003	0.430±0.003
声学未匿名化梅尔频谱	**0.517±0.004**	**0.493±0.002**	**0.466±0.003**	**0.455±0.002**

表16. 音频扩展实验中，场景（a）：自行车组装和场景（b）：3D打印机组装的开放词汇描述结果，包括有空类和无空类时的余弦相似度值。

模态	场景（a）		场景（b）	
	余弦相似度（↑）			
	不含空类	含空类	不含空类	含空类
声学匿名化编码器	0.361±0.030	0.341±0.018	0.316±0.003	0.323±0.004
声学非匿名化编码器	0.375±0.001	0.354±0.001	0.328±0.001	0.338±0.002
声学匿名化梅尔频谱	0.364±0.003	0.348±0.002	0.326±0.002	0.322±0.001
声学未匿名化梅尔频谱	**0.381±0.002**	**0.359±0.002**	**0.330±0.001**	**0.340±0.001**

表17。场景（a）：自行车组装和场景（b）：3D打印机组装的跨模态对齐结果，包括召回率@5、召回率@1和Top-1指标。

模态	场景（a）			场景（b）		
	召回率@1 (↑)	召回率@5 (↑)	Top-1 (↑)	召回率@1 (↑)	召回率@5（↑）	Top-1（↑）
声学 + 文本（匿名化编码器）	0.241±0.014	0.583±0.025	0.342±0.016	0.227±0.013	0.567±0.022	0.329±0.015
声学 + 文本（非匿名化编码器）	0.251± 0.001	0.605±0.002	0.355±0.002	0.237±0.001	0.589±0.001	0.341 ±0.001
声学 + 文本（匿名化梅尔频谱）	0.246 ± 0.001	0.595 ±0.002	0.350±0.001	0.233±0.001	0.578± 0.002	0.336 ±0.001
声学 + 文本（非匿名化梅尔频谱）	**0.254±0.001**	**0.613 ±0.002**	**0.360±0.001**	**0.238±0.001**	**0.597±0.001**	**0.345±0.001**

表18。场景（a）自行车组装和场景（b）3D打印机组装中工具接触/撞击检测的宏平均F1分数，分别报告包含和不包含空类的结果。与通用人类活动识别不同，音频有助于检测工具接触事件：在 V+I 中加入声学数据后，场景（a）的 F1 从 0.905 提升至 0.923，场景（b）则从 0.898 提升至 0.914。

模态	场景（a）		场景（b）	
	宏平均F1（↑）			
	无空类	Null	无空类	Null
惯性(I)	0.872±0.008	0.847±0.009	0.861±0.006	0.836±0.006
声学(A)	0.783±0.012	0.762±0.013	0.772±0.005	0.750±0.006
视觉(V)	0.841±0.009	0.817±0.010	0.833±0.004	0.808±0.005
I +A	0.898±0.007	0.871±0.008	0.889±0.004	0.863±0.004
A + 伏	0.863±0.009	0.838±0.010	0.855±0.004	0.830±0.005
I +伏	0.905±0.006	0.879±0.007	0.898±0.003	0.872±0.004
I +安 +伏	**0.923±0.005**	**0.897±0.006**	**0.914±0.003**	**0.888±0.004**

一个工件。

14. 传感器设计依据

如表中所总结。19，Open- Marcie 中的感知模态是基于物理互补性而非性能最大化进行选择的。

惯性测量单元是主要的可穿戴运动模态，能够捕捉细粒度的手部和身体动态，这对于区分操作模式至关重要；然而，惯性积分容易产生漂移，因此需要互补信号——磁力计有助于稳定姿态并减少长期旋转漂移。气压计提供与姿势转换相关的垂直运动线索（例如站立、跪姿、弯腰），并提供一种与高度相关的独立信号。

尽管当前基准测试中未执行显式的传感器融合，但气压计被部署在身体的多个位置，以支持潜在的相对高度推理。在这种配置下，压力变化的突然改变会以相似的（共模）方式影响所有传感器，从而使差分压力测量能够提供更稳定的垂直位移线索。此外，还会记录环境温度，以通过气压（测高）关系确保物理一致的高度估计；该关系描述了气压如何随高度变化 [4]。在此背景下，温度充当辅助环境变量，用于支持有依据的垂直运动解读，而不是作为独立的活动识别信号。

第一人称视角 RGB-D 感知通过捕捉物体身份、空间关系和手物交互提供语义依据，而外部视角多视图 RGB-D 摄像机则提供全局场景上下文，以及用于标注和跨模态对齐的稳定参考。激光雷达深度可提升不同光照

表19。OpenMarcie 中传感器模态及其预期作用。

模态	物理信号	与装配相关的属性	贡献
惯性测量单元（手腕、头部）	加速度、角速度、方向	工具操作、时间分段	High
磁力计	磁场方向	航向稳定、方向一致性	辅助
气压计（多位置）	气压（百帕）	姿势转换、垂直位移线索	中
温度（环境）	环境温度	环境背景，气压公式	低（辅助）
光谱仪	材料光谱反射率	材料区分（金属与塑料）	低（探索性）
热成像相机	表面温度分布	接触时长、摩擦热	低（探索性）
第一人称 RGB-D	视觉外观 +深度	物体身份，手-物体交互	High
第三人称 RGB-D（多视角）	场景级 RGB-D	全局空间上下文	High
立体声	乐器声/环境声	工具-材料交互特征	中
激光雷达深度	主动深度感知	三维空间结构、物体距离	高（结合视觉）

条件下的空间鲁棒性，并增强三维推理能力。

立体声音频能够捕捉工具与材料相互作用的特征，补充运动和视觉信息。尽管单独使用声学信息进行动作识别的表现有限，但它对于检测工具接触事件十分有用，如表中所示。18。

最后，基于光谱仪和热成像相机的视觉传感器也作为探索性模态纳入其中，以支持未来关于材料感知交互建模，以及捕捉人眼可见光谱之外物理信号的研究。

总体而言，这种多模态设计体现了一种有条理的尝试，旨在捕捉工业活动中互补的物理方面，而非一味增加模态数量。

15. 现代架构

表20和21 总结了现代多模态架构在场景（a）和场景（b）中的人类活动识别（HAR）与开放词汇图像描述生成任务上的表现。在 HAR 任务中，受 Perceiver IO 启发的模型 [20]始终优于后期融合基线；利用所有模态时效果最佳，在场景（a）中的宏平均F1分别达到 0.915（无空类）和 0.886（空类），在场景（b）中则分别达到 0.842（无空类）和 0.746（空类）。同样，在开放词汇图像描述生成任务中，基于大语言模型的 AnyMAL [22] 方法优于后期融合，尤其是在纳入所有模态时表现突出：场景（a）的余弦相似度分数为 0.643（无空类）和 0.610（空类），场景（b）则为 0.741（无空类）和 0.740（空类）。总体而言，结果表明，现代多模态融合架构和基于大语言模型的方法始终优于后期融合基线，尤其是在使用所有模态时。

16. 未来研究方向

除了本研究提出的基准测试外，我们的数据集还为研究界开辟了几个颇具前景的方向：

表 20. 在场景（a）和场景（b）中，包含和不包含空类时的人类活动识别宏平均F1分数。结果按后期融合基线以及使用所有模态（包括惯性测量单元、音频、视频、气压计、温度、光谱仪和热信号）的 Perceiver IO 基线分组。

模态	场景（a）		场景（b）	
	宏平均F1 (↑)			
	无空类	Null	无空类	Null
晚期融合				
I +A + V	0.859±0.010	0.831±0.011	0.763±0.003	0.676±0.003
All	0.891±0.008	0.862±0.008	0.825±0.003	0.731±0.004
Perceiver IO				
I+A+V	0.882±0.007	0.853±0.008	0.779±0.003	0.690±0.003
All	**0.915±0.005**	**0.886±0.006**	**0.842±0.002**	**0.746±0.003**

表 21. 场景（a）和场景（b）在有无空类情况下的开放词汇描述结果。结果按后期融合基线和 AnyMAL 基于大语言模型的基线分组。

模态	场景（a）		场景（b）	
	余弦相似度 (↑)			
	无空类	空类	无空类	Null
晚期融合				
I +A + V	0.547±0.020	0.519±0.017	0.647±0.001	0.646±0.003
All	0.593±0.015	0.563±0.014	0.691±0.002	0.690±0.002
AnyMAL				
I+A+V	0.622±0.012	0.562±0.012	0.714±0.002	0.693±0.002
All	**0.643±0.010**	**0.610±0.010**	**0.741±0.002**	**0.740±0.002**

1. **程序规划与任务分解** 对长期工业工作流程的层级结构进行建模，使系统能够学习任务图并分割复杂序列。未来的工作可以探索推断工作流依赖关系、优化分解或跨任务变体进行泛化的方法。

2. **技能评估与专业能力建模** 参与者差异为建模熟练度、效率和学习进程提供了机会。例如，未来方向包括自动技能分类、对智能体之间的专业能力迁移进行建模，以及根据行为信号设计自适应训练干预。

3. **意图预测与早期行动预测**根据部分多模态

模态观测预判即将发生的动作或目标，对于主动辅助与协作至关重要。潜在研究方向包括用于早期预测的多模态融合策略、部分观测序列中的目标推断，以及实时辅助系统。

4. **细粒度动作分割与角色理解**重叠与并发动作为分割和角色推断带来了颇具挑战性的测试平台。尚待解决的问题包括对多标签时间边界进行建模、学习多智能体或多阶段任务中的角色动态，以及将分割与下游规划联系起来。

5. **姿态估计与肢体语言推理**借助同步 RGB-D 和惯性数据，未来的研究可以推进全身姿态估计、活动条件化姿态预测，以及真实工业场景中的非语言意图识别。

6. **跨模态知识迁移**该数据集支持跨模态迁移学习，例如利用视觉或语言监督惯性或声学模型。这对于隐私敏感或传感器受限场景尤为重要。值得探索的方向包括跨模态蒸馏、模态丢弃鲁棒性和无监督对齐。

7. **跨模态生成与仿真**从一种模态生成另一种模态（例如根据指令生成 IMU 轨迹，或重建缺失的视频），能够支持稳健的模仿学习和仿真。未来的研究可以探索生成建模、仿真到现实迁移，以及面向工业任务的合成数据增强。