PhysLab：物理实验多粒度视觉解析的基准数据集

Minghao Zou
Shandong University of Science and Technology
Qingdao, China

Qingtian Zeng
Shandong University of Science and Technology
Qingdao, China

Yongping Miao
Shandong University of Science and Technology
Qingdao, China

Shangkun Liu
Shandong University of Science and Technology
Qingdao, China

Zilong Wang
Shandong University of Science and Technology
Qingdao, China

Hantao Liu
Cardiff University
Cardiff, UK

Wei Zhou
Cardiff University
Cardiff, UK

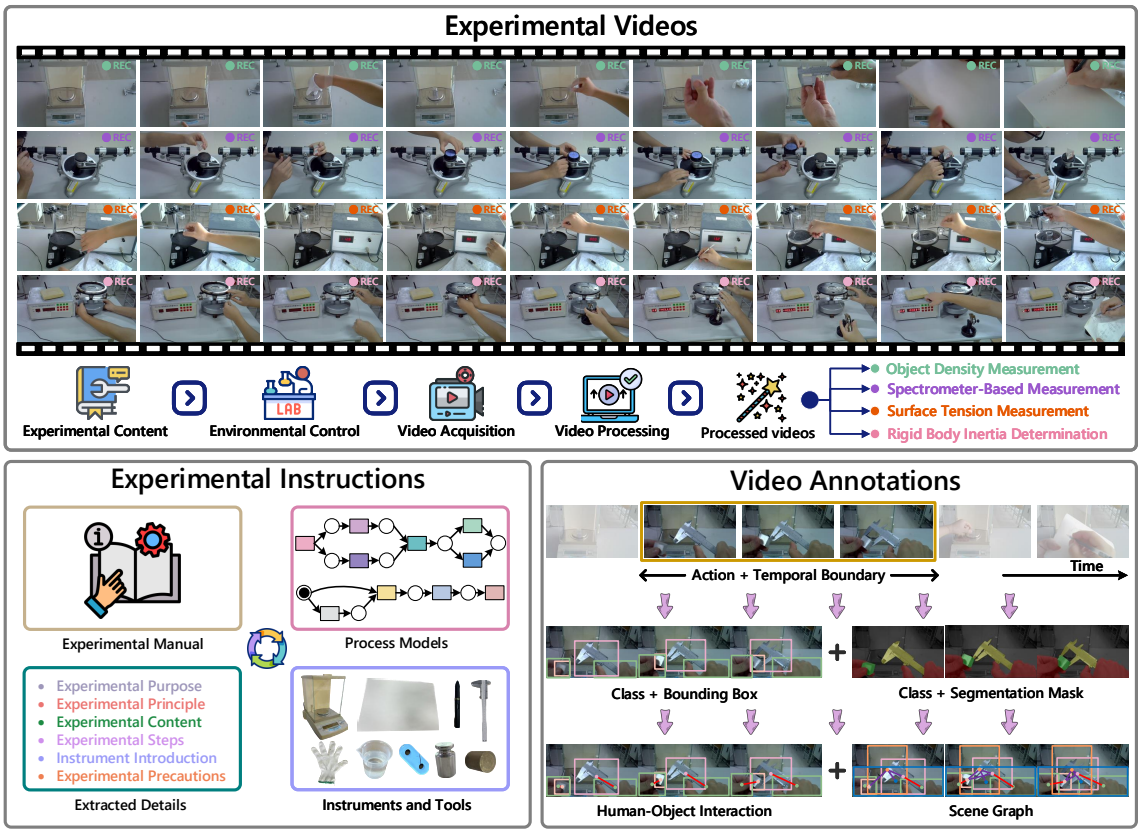

Figure 1: 我们提出的 PhysLab 数据集内容概述。PhysLab 数据集专注于在复杂的物理实验室环境中进行的实验任务。1）顶部行展示了数据收集过程和代表性数据样本。2）第二行展示了与物理实验相关的独特操作说明，包括详细程序、仪器使用和 Petri 网格式的流程模型（左下），以及多粒度注释（右下）。

Abstract

图像和视频的视觉解析对于广泛的实际应用至关重要。然而，这一领域的进展受到现有数据集限制的约束：(1) 注释的细粒度不足，妨碍了细致的场景理解和高层次的推理；(2) 领域的覆盖范围有限，尤其是缺乏适合教育场景的数据集；以及 (3) 缺乏明确的程序性指导，具有最小的逻辑规则且对结构化任务过程的表现不足。为了解决这些缺口，我们引入了 PhysLab，这是首个捕捉学生进行复杂物理实验的视频数据集。该数据集包含四个具有代表性的实验，特点是多样的科学仪器和丰富的人物-物体交互（HOI）模式。PhysLab 由 620 个长视频组成，提供多级注释，支持包括动作识别、物体检测、HOI 分析等在内的各种视觉任务。我们建立了强有力的基准并进行了广泛的评估，以强调程序性教育视频解析中的关键挑战。我们期望 PhysLab 能够成为推进细粒度视觉解析的重要资源，促进智能课堂系统，并推动计算机视觉与教育技术之间的更紧密结合。数据集和评估工具包可在 https://github.com/ZMH-SDUST/PhysLab 公开获取。

www.xueshuxiangzi.com

Keywords

Procedural Video, Physics Lab Education, Visual Parsing, Multi-Granularity Annotation

1 引言

视觉解析是计算机视觉中的基础研究，涵盖了包括动作识别、目标检测和人机交互检测 [53] 在内的广泛任务。在实际应用中，如监控、自动驾驶和工业检测，它起到了关键作用。近年来，由于深度学习的进步和大规模标注数据集 [61] 的可用性，该领域取得了显著的进展。然而，这种以数据为中心的范式也暴露了对高质量、多样化标注的强烈依赖。尽管现有数据集规模较大，但在支持多任务协同建模和丰富语义理解方面仍显不足 [24] 。具体来说，它们往往针对单任务学习或仅提供粗糙的标签，缺乏联合建模动作、对象和交互所需的结构化细粒度标注 [18] 。此外，大多数数据集主要围绕日常场景构建，如家庭 [46] 、街道 [29] 和厨房 [22] ，而忽略了诸如教育等具认知需求的领域。结果，许多基线模型倾向于优先处理静态外观线索或短期视觉模式，在建模复杂程序任务固有的因果关系、时间依赖性及物理约束方面表现不足 [51] 。这些限制阻碍了具备更深层次、人类般理解能力模型的发展

。

与传统视觉场景相比，物理实验展示了显著更高的时空复杂性和认知要求 [21] 。实验过程通常涉及多步骤操作、仪器配置和参数调整，所有这些都受严格的任务依赖性和程序顺序控制 [64] 。这些任务还具有频繁且详细的人机互动，包括工具切换和反馈观察，这需要细粒度的识别和跨模态推理 [16] 。此外，这些过程常常由隐含的物理知识和因果动态支持，光靠低层视觉特征无法有效捕获 [38] 。为了促进对这种复杂且认知丰富场景的研究，我们引入了 PhysLab，这是首个专为物理实验指导而设计的多任务视频数据集。

如图 1 所示，PhysLab 专注于四个具代表性的本科物理实验，涵盖多种科学仪器和典型学生行为。该数据集包含 620 个长时视频（总计 31 小时），每个视频记录了真实的学生实验过程。视频被分割为动作片段，并丰富地标注了层次标签，包括动作类别、物体位置、互动关系和操作序列。这些标注为多颗粒度和多任务学习提供了统一平台。此外，数据集还结合了结构化元数据，编码实验类型、步骤流程和所用工具，这有助于进行较高层次的任务如过程建模和实验状态推理。在这种背景下，过程模型指的是对实验工作流程进行结构化表示，捕捉操作之间的逻辑和时间约束。它帮助理解实验怎样随时间展开，并支持在任务执行期间对进展、正确性和偏差进行推理。

为了系统地评估 PhysLab 带来的挑战，我们进行了广泛的基准实验，涵盖了 8 个主流的视觉解析任务。评估结果突显了数据集的复杂性，并揭示了当前方法的重大局限性，特别是在视觉上下文融合不足、缺乏先验知识以及明显的类间性能差异方面。我们相信，PhysLab 为推动复杂行为建模的研究、改进教育过程分析以及弥合视觉理解与智能教育系统之间的差距提供了一个有价值的基础。

2 相关工作

2.1 时间解析数据集

时间解析数据集专注于对视频的时间动态进行建模，并支持诸如动作分类、时间动作定位和动作分割等任务 [44, 47] 。早期的数据集如 UCF101 [43] 、HMDB51 [20] 和 Kinetics [4] 包含各种描述日常人类动作的短、编辑过的片段，并已被广泛用于动作分类任务。然而，这些数据集在动作类型上提供的多样性有限，并且缺乏细粒度的时间结构。为了解决这一问题，后续的数据集如 THUMOS14 [17] 、ActivityNet1.3 [3] 和 FineAction [30] 引入了未剪裁的具有更密集时间注释的视频，支持动作检测和分割方面的研究

。

最近，随着社区向建模结构化活动转变，已经出现了几个程序性视频数据集。数据集如 Breakfast [19] 、COIN [46] 和 Assembly101 [40] 专注于具有明确目标的多步骤指导任务，支持程序理解的研究。此外，像 EgoPER [22] 、CaptainCook4D [36] 和 IndustReal [39] 这样的数据集结合了执行错误和偏差的注释，促进了任务工作流中异常检测的研究。尽管最近有所进展，但现有数据集仍然主要局限于烹饪和机械装配 [45] 等狭窄领域。在学生实验室实验资源方面存在显著不足，这对于推进视觉感知与智能教育系统交叉研究至关重要 [37] 。

2.2 空间解析数据集

相比之下，空间解析数据集关注于静态帧中的对象识别和关系推理。像 COCO [28] 和 Objects365 [41] 这样的通用对象检测数据集覆盖了广泛的日常场景，而特定领域的数据集—例如用于自动驾驶的 TJU-DHD [34] 、用于工业检测的 MVTec AD [2] 和 Read-IAD [48] ，以及用于医疗设备的 HIOD [14] —提供了高质量的样本，以应对诸如遮挡、光照变化和尺度多样性等挑战。在对象检测的基础上，包括 ADE20K [59] 和 COCONut [10] 在内的实例分割数据集提供了像素级别的注释，使更细粒度的视觉理解成为可能。

超越了对象级别的任务，用于 HOI 检测和场景图生成的数据集旨在捕捉实体之间的语义关系，为高级视觉推理奠定基础。代表性的数据集如 HICO-DET、V-COCO、Visual Genome 和 Open Images V4/V6 提供了全面的人物-物体和物体-物体交互注释。尽管取得了这些进展，大多数现有数据集缺乏多粒度行为注释和任务级别的语义对齐，这限制了它们在复杂程序建模和因果行为推断中的有效性。相比之下，PhysLab 支持时间和空间解析，提供丰富的注释，使从细粒度时间动作分割到帧级关系推理等广泛的任务成为可能。

3 PhysLab 数据集

我们选择了四个具有代表性的大学物理实验作为数据收集的基础，涵盖了以下实验任务：基于光谱仪的角度测量、刚体转动惯量测定、表面张力测量和物体密度测量。为了确保数据集的真实性和研究价值，在数据采集过程中考虑了以下因素：

- 任务和行为多样性。所选实验涵盖了各种任务、工具和操作模式，涉及多样的操作程序和人机交互。
- 环境多样性。视频录制在多种条件下进行，包括多个摄像机角度（三个不同角度）、焦距、灯光设置、时间和实验室环境的变化。这种多样性增强了数据集在实际部署场景中的鲁棒性和普适性 [35] 。
- 标准化实验协议。所有程序严格遵循官方大学物理实验手册，该手册提供了目标、设备规格、程序步骤、数据记录要求和安全指南的详细描述。这确保了程序的一致性、可重复性和教学的有效性。
- 学生自主性和执行灵活性。每个实验都是由学生独立进行的，没有外部干预，从而允许在执行顺序、动作持续时间和工具操作上的自然变化。该设置捕捉成功的操作和程序错误，支持下游任务，如异常检测、程序遵从性分析和学习行为建模 [22] 。

www.xueshuxiangzi.com

用于视频采集，我们采用了 Newmine Q40 高清摄像机，该摄像机支持 4K 分辨率（3840×2160）、帧率 30 FPS，并配有 90 度无失真镜头。该设备提供了出色的图像清晰度、宽广的视野覆盖和稳定的长时间性能，非常适合在实验室设置中连续部署和自动数据采集场景。

经过严格的筛选和清理过程，我们汇总了总计 620 个实验视频，每个视频的时长从 1 到 8 分钟不等。整个数据集包含大约 31 小时的注释视频内容，形成了一个高质量且多样化的程序物理实验语料库。该资源非常适合用于视觉解析和实验过程建模。

3.1 数据标注

我们与物理实验中心合作，建立了一支专业的标注团队，并采用了“交叉标注 & 多轮验证”协议 [52]，以确保高质量和一致性的标注。为了支持多粒度的视觉解析任务，我们进行了时间和空间标注。标注的类别模式在图 2 中示意，详细的定量统计数据在我们开源网站上提供。

时间注释。对于每个实验视频，我们沿时间轴注释了每个操作步骤的类别、开始时间和结束时间。这些注释有助于分段视频中的动作分类和未剪辑视频中的动作定位 [11]。所有标签严格遵循官方实验手册中概述的标准化程序，确保数据集中语义的一致性。为了提高注释的客观性——特别是在划分精确的时间边界时——每个视频由两位经验丰富的注释员使用 ELAN [50] 工具独立标注。ELAN 的多层时间轴结构可以对复杂的程序性行为进行详细的、分层的描述。在独立注释之后，结果被合并并手动调和，以解决差异并确保时间一致性。总共注释了 3,873 个动作实例，每个实例的平均持续时间约为 20 秒，涵盖 32 种不同类型的实验步骤。这些全面的注释为程序性行为分析提供了坚实的基础。

空间标注。为了支持细粒度的空间解析，我们从实验视频中提取关键帧，并利用图像相似性过滤技术消除冗余帧，从而获得了一个包含 4,500 个高质量帧的多样化子集。在这些帧上，我们为 34 种实验仪器和物体标注了边界框，这些类别是根据实验手册中的官方设备规格定义的。此外，使用了 24 种交互动词来描述操作员与仪器之间的功能关系，基于实验动作的语义。因此，每个标注的帧被结构化为一个形式为 < 操作者，交互动词，仪器 > 的三元组。该标注设计支持 HOI 检测任务 [54]，并支持构建用于高级推理和过程理解的交互图 [62]。鉴于物体的高密度、频繁的交互以及物理实验场景中典型的标注复杂性，一个由 5 名研究生和 18 名本科生组成的专门标注团队使用 Labelme 工具进行了标注过程。为了保证标注质量，我们实施了多轮交叉检查和基于样本的审查程序 [60]。

3.2 数据统计

随着视觉解析技术的进步，研究数据集已经从捕捉简单语义场景逐渐演变为包含更复杂和认知要求更高的程序任务视频 [42]。表 1 总结并比较了具有代表性的程序视频数据集，突出显示了本研究中开发的 PhysLab 数据集的优势和独特特征

。

早期的数据集如 Breakfast [19] 和 CrossTask [63] 主要专注于日常活动。虽然这些数据集采集成本相对较低，但其提供的过程性内容常常过于简单，缺乏现实世界任务执行中所具有的灵活性、多样性和易出错的行为。因此，它们在建模复杂任务动态或分析执行不确定性方面的适用性仍然有限。

最近，研究已经转向工业和模拟装配场景，这在诸如 Assembly101 [40]、HA4M [8]、ATTACH [1] 和 IndustReal [39] 等数据集中有所体现。这些数据集具有更复杂的任务流程，通常涉及家具、玩具或工业设备的装配。它们包括时间和空间层面的注释，大大增强了其描述能力并支持细粒度分析。在这一趋势的基础上，我们提出的 PhysLab 数据集是首个以智能教育领域为目标的数据集。它包含 620 个由大学生进行的真实物理实验视频，总计约 31 小时的录像，使其在规模上跻身于高端程序视频数据集之列。更重要的是，PhysLab 引入了以下独特的特征：

- 高任务真实性。该数据集直接收集自真实的实验环境，捕捉了真实的学生行为，包括错误、不正确的操作顺序和任务偏差。这种真实性为实验性学习过程的自然变异性提供了宝贵的见解。
- 高过程灵活性。不同的学生在进行相同实验时通常会采用不同的操作顺序和执行风格。这种多样性促进了灵活和动态过程结构的建模，这对于行为理解和过程概括的研究特别有益。
- 丰富的注释粒度。除了动作的精确时间边界和物体级别的空间标签之外，数据集还包括 24 类交互动词。这些动词用于构建结构化的 HOI 三元组，以捕捉学生和工具之间详细的交互动态。这些多粒度注释极大地提升了数据集在多任务建模中的价值，如动作识别、HOI 检测和过程挖掘。

总之，PhysLab 以其实验程序的全面覆盖、行为复杂性和细粒度注释脱颖而出，使其成为高质量、多用途的数据集，非常适合用于实验过程监控和视觉过程分析的下游应用。

4 选定任务

PhysLab 数据集具有多样且多粒度的标注，支持多种视觉分析任务，包括时序动作定位、动作分类、动作识别、目标检测、遮挡检测、实例分割、HOI 检测和场景图生成。在本节中，我们评估了几种代表性方法在 PhysLab 上的表现，重点关注两个关键任务：动作识别和 HOI 检测。这些任务用于从时间和空间的角度，检验物理实验场景中视觉解析的独特挑战。有关其他任务和基准，请参阅官方开源存储库。

4.1 动作识别

任务表述。动作识别的目标是从未剪辑的视频中定位和分类动作。在程序性任务分析中，这一任务通常被分为两个子任务 [23, 31]：

- 动作对齐。在这个任务中，测试视频的转录（即动作标签序列）是已知的，模型需要预测每个动作片段的时间边界，以与真实值（GT）对齐。
- 动作分割。在这个更具挑战性的任务中，没有可用的文本记录。模型必须同时分割动作的时间边界，并准确预测相应的动作标签。

评估指标。我们采用两个标准指标：帧均值（MoF）和交并比（IoU）[64]。MoF 衡量预测动作标签与 GT 匹配的视频帧的百分比。IoU 量化每个动作的预测边界集与 GT 边界集之间的平均重叠。两个指标分别在公式（1）和公式（2）中定义。

$$\frac{1}{T}\sum_{i=1}^{T}(\hat{y}_i = y_i) \tag{1}$$

$$\frac{1}{A}\sum_{a}|G_a \cap D_a| \,/|G_a \cup D_a| \tag{2}$$

www.xueshuxiangzi.com

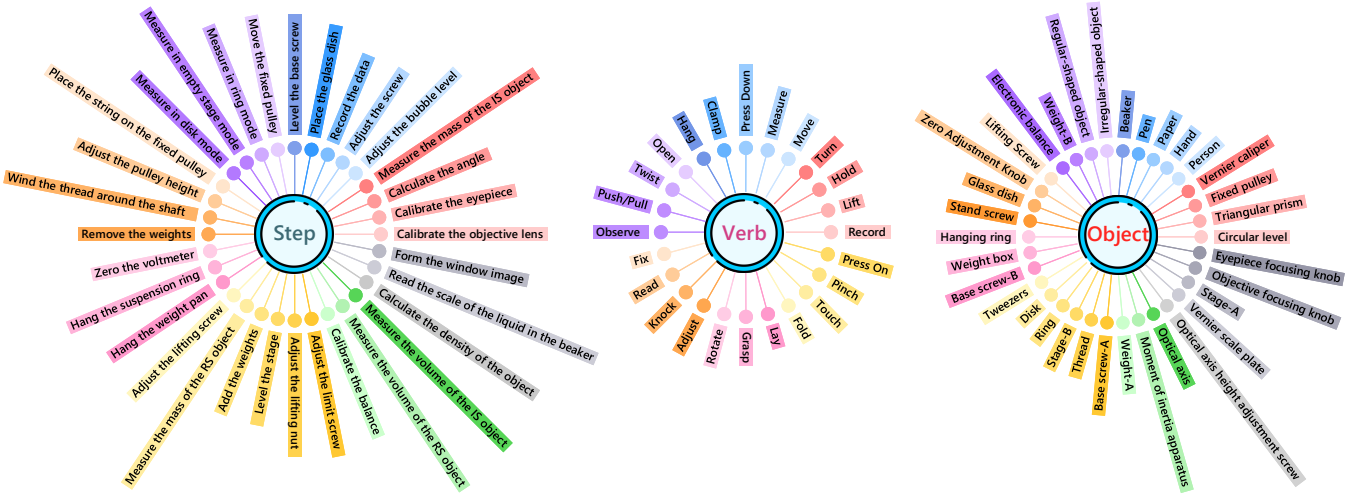

Figure 2: PhysLab 数据集中实验步骤、交互动词和对象的分类法。

Table 1: 程序视觉解析视频数据集的比较。AR：动作识别，OD：物体检测，IS：实例分割，HOID：人-物体交互检测，Flex.：任务的执行顺序不止一种，PEs：程序错误。

Dataset	Year	Domain/Environment	Tasks				Procedural complexity		Dataset Size	
			AR	OD	IS	HOID	Flex.	PEs	Videos	Hours
Breakfast [19]	2014	Cooking	✓	✗	✗	✗	✗	✗	1.7K	77.0
Epic-Kitchens [9]	2018	Cooking	✓	✗	✗	✗	✗	✗	432	55.0
CorssTask [63]	2019	Daily Life	✓	✗	✗	✗	✗	✗	4.7K	375.0
COIN [46]	2019	Daily Life	✓	✗	✗	✗	✗	✗	11.8K	476.0
Assembly101 [40]	2022	Toy Assembly	✓	✗	✗	✗	✓	✓	1.0K	167.0
HA4M [8]	2022	Industrial Assembly	✓	✗	✗	✗	✓	✗	217	5.9
ATTACH [1]	2023	Furniture Assembly	✓	✗	✗	✗	✓	✗	378	17.2
ATA [12]	2023	Toy Assembly	✓	✓	✗	✗	✓	✗	1.2K	24.8
IndustReal [39]	2024	Toy Assembly	✓	✓	✗	✗	✓	✓	84	5.8
EgoPER [22]	2024	Cooking	✓	✓	✗	✗	✓	✗	386	28.0
PhysLab (Ours)	2025	Physics Experiment	✓	✓	✓	✓	✓	✓	620	31.0

，其中 $\hat{y}_i$ 和 y_i 分别表示帧 i 的预测和 GT 动作标签，T 是视频中的总帧数。G_a 和 D_a 分别表示标记为动作 a 的 GT 和预测帧集合，A 是不同动作类别的总数。

结果。我们在三个数据集——PhysLab、Breakfast 和 CrossTask——上评估了四种具有代表性的方法，涵盖了动作对齐和动作分割任务。如表 2 所示，这些方法在 PhysLab 上的整体表现不如其他两个数据集，尤其是在 MoF 方面。这一性能差距可归因于 PhysLab 的几个固有因素：大量的细粒度实验动作的存在、频繁的遮挡、复杂的人物与物体交互，以及不同操作之间的高视觉相似性。这些特性大大增加了帧级动作分类和时间边界定位的难度。

此外，PhysLab 展现出更强的区分模型性能的能力。在动作对齐任务中，PhysLab 上表现最佳和最差的模型之间的 IoU 差距达到 21.2，而在 Breakfast 和 CrossTask 上的相应差距仅为 5.0 和 3.1。类似的趋势也出现在动作分割任务上，其中 PhysLab 的 IoU 差距为 20.3，明显大于其他数据集。这些结果表明，PhysLab 的复杂视觉和过程特征不仅对动作识别提出了更大的挑战，还提供了一个更严格的基准来评估模型的鲁棒性和泛化能力。

Table 2: 在 PhysLab、Breakfast 和 CrossTask 上的动作对齐和动作分割性能比较。所有报告的值均乘以 100 以提高可读性。

Task	Method	PhysLab		Breakfast		CrossTask	
		MoF	IoU	MoF	IoU	MoF	IoU
Alignment	CDFL [23]	22.8	8.0	63.0	45.8	46.7	17.2
	TASL [31]	30.5	20.4	65.8	49.9	57.1	19.1
	POC [32]	36.0	28.0	56.1	46.7	53.3	18.9
	AL-PKD [64]	41.5	29.2	67.6	50.8	62.7	20.3
Segmentation	CDFL [23]	12.0	5.7	50.2	33.7	32.5	11.8
	TASL [31]	28.7	19.7	49.9	36.6	42.7	14.9
	POC [32]	32.2	25.1	47.1	39.4	44.1	16.3
	AL-PKD [64]	38.7	26.0	51.5	39.7	44.7	18.6

www.xueshuxiangzi.com

Table 3: 人体-物体交互检测在 PhysLab 和 HICO-DET 上的性能比较。所有报告的数值都乘以 100 以增加可读性。

		PhysLab			HICO-DET		
Method	Backbone	Full	Rare	Non-Rare	Full	Rare	Non-Rare
STIP [58]	ResNet-50	62.62	61.00	62.73	32.22	28.15	33.43
GEN-VLKT $_S$ [26]	ResNet-50	58.71	75.00	57.58	33.75	29.25	35.10
OCN [55]	ResNet-50	52.19	68.01	51.10	30.91	25.26	32.51
LOGICHOI [25]	ResNet-50	53.34	50.97	62.50	35.47	32.03	36.22
TED-Net [49]	ResNet-50	60.97	69.89	60.35	34.00	29.88	35.24
SOV-STG-S [7]	ResNet-50	45.42	34.17	49.50	33.80	29.28	35.15
OCN [55]	ResNet-101	49.50	50.00	49.46	31.43	25.80	33.11
UPT [56]	ResNet-101	65.28	63.47	65.43	32.31	28.55	33.44
GEN-VLKT $_L$ [26]	ResNet-101	60.19	64.25	69.91	34.95	31.18	36.08
SOV-STG-L [7]	ResNet-101	52.35	46.59	58.44	35.01	30.63	36.32
ERNet [27]	ENetV2-XL	56.39	49.26	62.36	35.92	30.13	38.29
PViC [57]	Swin-L	71.42	68.47	71.62	44.32	44.61	44.24
FGAHOI [33]	Swin-L	60.79	54.89	67.23	37.18	20.71	39.11
SOV-STG-Swin-L [7]	Swin-L	56.48	45.23	58.44	43.35	42.25	43.69

4.2 人体-物体交互检测

任务表述。该任务涉及从图像中识别交互三元组，每个三元组由人的主体的边界框、交互对象的类别和边界框，以及描述它们关系的交互动词组成 [13] 。

评估指标。我们采用完整、稀有和非稀有设置下的标准平均精度（mAP），遵循 HICO-DET 评估协议 [6] 。如果且仅当满足以下所有条件时，预测的交互三元组被认为是真阳性（TP）：

(1) 预测的人和物体的类别标签是正确的。

(2) 预测的人和物体框与对应的 GT 框的 IoU（空间级别）各自大于 0.5。

(3) 预测的交互动词与 GT 标签匹配。

根据这些标准，计算每个交互类别的平均精度（AP），并定义整体 mAP 为：

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{3}$$

，其中 C 是交互类别的数量，AP_i 表示类别 i 的平均精度。

结果。我们在 PhysLab 和 HICO-DET 上评估了 10 种具有代表性的 HOI 检测方法，结果汇总在表 3 中。总体而言，模型在 PhysLab 上的性能通常高于 HICO-DET，主要是因为 PhysLab 专注于物理实验中结构化和特定领域的交互，而 HICO-DET 涵盖了多样化和开放的场景。尽管有这种性能优势，PhysLab 提出了两个独特的挑战。首先，它明显的类别不平衡导致了类别间显著的性能变化。与 HICO-DET 不同，HICO-DET 上的大多数模型在非稀有类别上表现更好，而在 PhysLab 上的模型则表现出不同的趋势。例如，STIP [58] 和 LOGICHOI [25] 在非稀有交互上表现更好，而 OCN [55] 在稀有类别上取得了更优的性能（在稀有类别上获得了 68.01 的 mAP，对比非稀有类别的 51.10）。值得注意的是，GEN-VLKT $_S$ [26] 在两组之间的 mAP 差距达到了 17.42，强调了在不平衡标签分布下 PhysLab 所提出的增加的泛化需求。

其次，模型性能在每个评估子集中的不同模型之间表现出更大的波动。具体来说，在 PhysLab 的全面、稀有和非稀有子集内，最佳和最差模型之间的性能差距分别达到 57.2 % 、119.5 % 和 44.8 % ，这些数值明显高于在 HICO-DET 中观察到的对应差距，即 43.4 % 、115.4 % 和 20.3 % 。这表明，PhysLab 的交互情境不仅放大了不同模型架构的优缺点，还更好地揭示了它们对不同交互类型和类别频率的适应性。因此，PhysLab 中显著的类别不平衡和更大的性能差异使其成为一个更具挑战性的基准 [5] 。这些特性强调了它在促进开发更强大和细粒度的 HOI 检测算法方面的价值，尤其是在实验或教学环境中 [15] 。

5 结论与未来工作

在本文中，我们介绍了 PhysLab，这是一个丰富标注的基准数据集，旨在推进物理实验过程的多粒度视觉解析。PhysLab 在计算机视觉与科学教育的交集处填补了关键空白，通过在真实实验背景中嵌入视觉识别任务。通过模拟实验活动中复杂的程序结构，PhysLab 促进了针对教育环境的可解释和目标导向的场景理解系统的发展。我们相信，PhysLab 为教育 AI 研究和更广泛的视觉社区提供了一个新颖且有影响力的视角。

我们正在积极扩充数据集，通过收集和注释六个其他代表性物理实验的视频数据。同时，我们计划扩展 PhysLab 以涵盖化学和生物学的实验程序，从而增加领域多样性并支持跨学科的泛化。此外，为了实现更强大和具有上下文意识的实验过程建模，我们旨在整合多模态信号，例如音频、文本指令和传感器数据，并探索跨模态对齐和融合策略，以更好地捕捉实验任务的语义结构。所有版本的数据集，以及基准协议和评估结果，都会通过我们的开源平台持续维护和发布，以鼓励重复性并促进研究社区内的协作开发。

References

[1] Dustin Aganian, Benedict Stephan, Markus Eisenbach, Corinna Stretz, and Horst-Michael Gross. 2023. ATTACH dataset: Annotated two-handed assembly actions for human action understanding. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11367–11373.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE/CVF, Long Beach, 9592–9600.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. IEEE/CVF, Boston, 961–970.

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Honolulu, 6299–6308.

www.xueshuxiangzi.com

[5] Sixian Chan, Xianpeng Zeng, Xinhua Wang, Jie Hu, and Cong Bai. 2024. Auxiliary Feature Fusion and Noise Suppression for HOI Detection. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 10 (2024), 1–18.

[6] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, IEEE, Lake Tahoe, 381–389.

[7] Junwen Chen, Yingcheng Wang, and Keiji Yanai. 2025. Focusing on what to Decode and what to Train: SOV Decoding with Specific Target Guided DeNoising and Vision Language Advisor. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 9416–9425.

[8] Grazia Cicirelli, Roberto Marani, Laura Romeo, Manuel García Domínguez, Jónathan Heras, Anna G Perri, and Tiziana D' Orazio. 2022. The HA4M dataset: Multi-Modal Monitoring of an assembly task for Human Action recognition in Manufacturing. *Scientific Data* 9, 1 (2022), 745.

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*. 720–736.

[10] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. 2024. Coconut: Modernizing coco segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Seattle, 21863–21873.

[11] Guodong Ding, Fadime Sener, and Angela Yao. 2023. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 2 (2023), 1011–1030.

[12] Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, and Behzad Dariush. 2023. Weakly-supervised action segmentation and unseen error detection in anomalous instructional videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10128–10138.

[13] Geng Han, Jiachen Zhao, Lele Zhang, and Fang Deng. 2024. A Survey of Human-Object Interaction Detection With Deep Learning. *IEEE Transactions on Emerging Topics in Computational Intelligence* 9 (2024), 3–26.

[14] Da Hu, Shuai Li, and Mengjun Wang. 2023. Object detection in hospital facilities: A comprehensive dataset and performance evaluation. *Engineering Applications of Artificial Intelligence* 123 (2023), 106223.

[15] Jiamian Hu, Hong Yuanyuan, Yihua Chen, He Wang, and Moriaki Yasuhara. 2024. Noisy Ostracods: A Fine-Grained, Imbalanced Real-World Dataset for Benchmarking Robust Machine Learning and Label Correction Methods. *Advances in Neural Information Processing Systems* 37 (2024), 50750–50771.

[16] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. 2024. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Seattle, 22072–22086.

[17] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding* 155 (2017), 1–23.

[18] Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *Comput. Surveys* 57, 10 (2025), 1–35.

[19] Hilde Kuehne, Ali Arslan, and Thomas Serre. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Columbus, 780–787.

[20] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*. IEEE, IEEE, Barcelona, 2556–2563.

[21] Ehsan Latif, Ramviyas Parasuraman, and Xiaoming Zhai. 2024. Physicsassistant: An LLM-powered interactive learning robot for physics lab investigations. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, IEEE, Pasadena, 864–871.

[22] Shih-Po Lee, Zijia Lu, Zekun Zhang, Minh Hoai, and Ehsan Elhamifar. 2024. Error detection in egocentric procedural task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Seattle, 18655–18666.

[23] Jun Li, Peng Lei, and Sinisa Todorovic. 2019. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6243–6251.

[24] Jianwei Li, Jun Xue, Rui Cao, Xiaoxia Du, Siyu Mo, Kehao Ran, and Zeyan Zhang. 2024. Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Seattle, 3184–3193.

[25] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. 2023. Neural-logic human-object interaction detection. *Advances in Neural Information Processing Systems* 36 (2023), 21158–21171.

[26] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20123–20132.

[27] JunYi Lim, Vishnu Monn Baskaran, Joanne Mun-Yee Lim, KokSheik Wong, John See, and Massimo Tistarelli. 2023. Ernet: An efficient and reliable human-object interaction detection network. *IEEE Transactions on Image Processing* 32 (2023), 964–979.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, Springer, Zurich, 740–755.

[29] Lihao Liu, Yanqi Cheng, Zhongying Deng, Shujun Wang, Dongdong Chen, Xiaowei Hu, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica Aviles-Rivero. 2024. TrafficMOT: A Challenging Dataset for Multi-Object Tracking in Complex Traffic Scenarios. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM New York, NY, Melbourne, 1265–1273.

[30] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. 2022. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing* 31 (2022), 6937–6950.

[31] Zijia Lu and Ehsan Elhamifar. 2021. Weakly-supervised action segmentation and alignment via transcript-aware union-of-subspaces learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8085–8095.

[32] Zijia Lu and Ehsan Elhamifar. 2022. Set-supervised action learning in procedural task videos via pairwise order consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19903–19913.

[33] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. 2024. Fgahoi: Fine-grained anchors for human-object interaction detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2024), 2415–2429.

[34] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. 2020. TJU-DHD: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing* 30 (2020), 207–219.

[35] Goran Paulin and Marina Ivasic-Kos. 2023. Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial intelligence review* 56, 9 (2023), 9221–9265.

[36] Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, et al. 2024. CaptainCook4D: A dataset for understanding errors in procedural activities. *Advances in Neural Information Processing Systems* 37 (2024), 135626–135679.

[37] Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. 2022. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour* 6, 9 (2022), 1257–1267.

[38] CFJ Pols and PJJM Dekkers. 2024. Redesigning a first year physics lab course on the basis of the procedural and conceptual knowledge in science model. *Physical Review Physics Education Research* 20, 1 (2024), 010117.

[39] Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons Van der Sommen, et al. 2024. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE/CVF, Waikoloa, 4365–4374.

[40] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, New Orleans, 21096–21106.

[41] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. IEEE/CVF, Seoul, 8430–8439.

[42] Yuhan Shen and Ehsan Elhamifar. 2024. Progress-aware online action segmentation for egocentric procedural task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18186–18197.

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012), 1–7.

[44] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence* 45, 3 (2022), 3200–3225.

[45] Hui Li Tan, Hongyuan Zhu, Joo-Hwee Lim, and Cheston Tan. 2021. A comprehensive survey of procedural video datasets. *Computer Vision and Image Understanding* 202 (2021), 103107.

[46] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Long Beach, 1207–1216.

www.xueshuxiangzi.com

[47] Binglu Wang, Yongqiang Zhao, Le Yang, Teng Long, and Xuelong Li. 2023. Temporal action localization in the deep learning era: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2023), 2171–2190.

[48] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. 2024. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Seattle, 22883–22892.

[49] Yuxiao Wang, Qi Liu, and Yu Lei. 2024. Ted-net: Dispersal attention for perceiving interaction region in indirectly-contact hoi detection. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 7 (2024), 5603–5615.

[50] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*. 1556–1559.

[51] Tao Wu, Runyu He, Gangshan Wu, and Limin Wang. 2024. Sportshhi: A dataset for human-human interaction detection in sports videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE/CVF, Seattle, 18537–18546.

[52] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 19368–19376.

[53] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. 2022. Mmpyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM international conference on multimedia*. ACM New York, NY, 6241–6249.

[54] Jiale Yu, Baopeng Zhang, Qirui Li, Haoyang Chen, and Zhu Teng. 2023. Hierarchical reasoning network with contrastive learning for few-shot human-object interaction recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4260–4268.

[55] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. 2022. Detecting human-object interactions with object-guided cross-modal calibrated semantics. In *Proceedings of the AAAI Conference on artificial intelligence*, Vol. 36. 3206–3214.

[56] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. 2022. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20104–20112.

[57] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. 2023. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10411–10421.

[58] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. 2022. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19548–19557.

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Honolulu, 633–641.

[60] Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, et al. 2025. BrowseComp-ZH: Benchmarking Web Browsing Ability of Large Language Models in Chinese. *arXiv preprint arXiv:2504.19314* (2025).

[61] Wei Zhou, Hadi Amirpour, Christian Timmerer, Guangtao Zhai, Patrick Le Callet, and Alan C Bovik. 2025. Perceptual Visual Quality Assessment: Principles, Methods, and Future Directions. *arXiv preprint arXiv:2503.00625* (2025), 1–6.

[62] Xuhan Zhu, Yifei Xing, Ruiping Wang, Yaowei Wang, and Xiangyuan Lan. 2024. Calibration for Long-tailed Scene Graph Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3037–3046.

[63] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, Long Beach, 3537–3545.

[64] Minghao Zou, Qingtian Zeng, and Xue Zhang. 2024. Weakly-supervised Action Learning in Procedural Task Videos via Process Knowledge Decomposition. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 7 (2024), 5575–5588.

www.xueshuxiangzi.com