# Towards Visual Discrimination and Reasoning of Real-World Physical Dynamics: Physics-Grounded Anomaly Detection

# 面向现实世界物理动力学的视觉判别与推理:基于物理的异常检测

Wenqiao Yao Xintao Chen Xiaohao Ming Xiaonan Yingna ShanghaiTech University University of Michigan, Ann Arbor Monash University

文乔 姚 陈新涛 肖浩 明 晓楠 英娜 上海科技大学 密歇根大学安娜堡分校 莫纳什大学

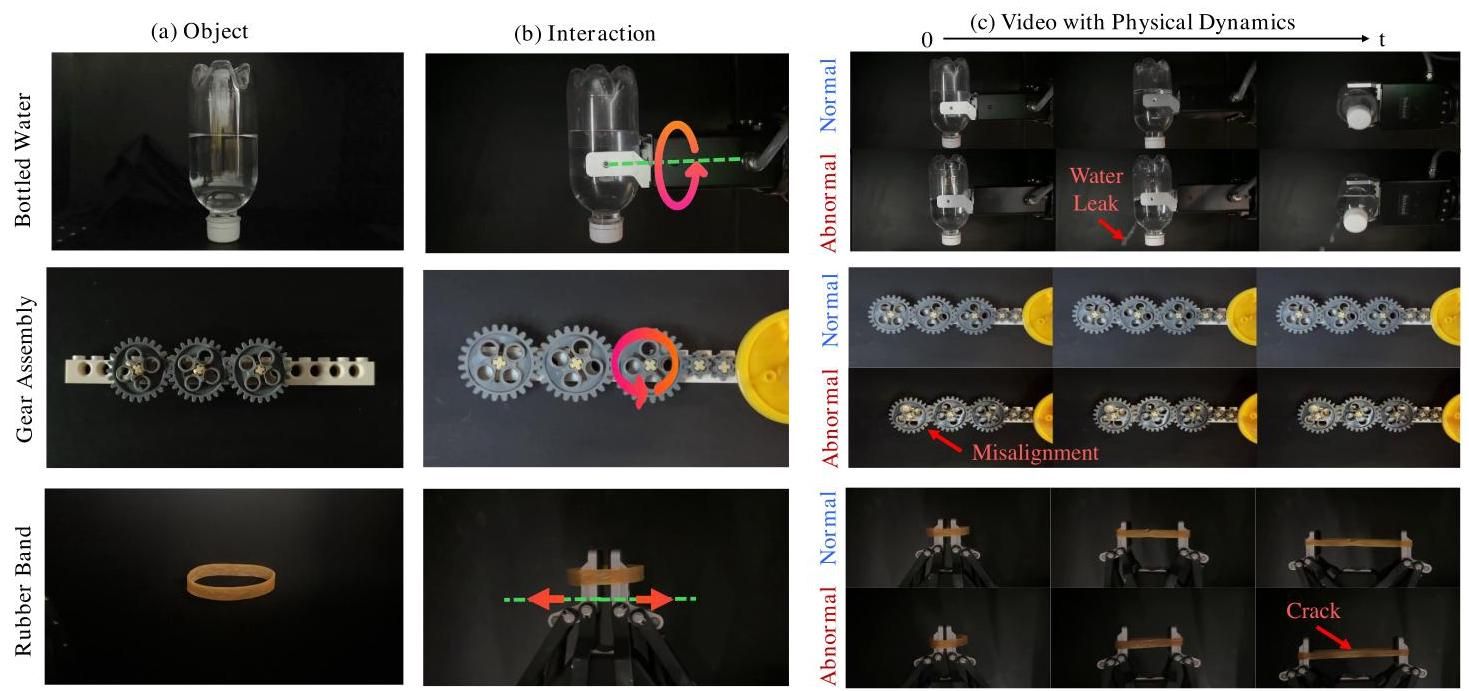


Figure 1. Towards visual discrimination of physical dynamics in real-world industrial object anomaly detection. We illustrate objects, interactions, and time-sequenced videos from the Physics-Grounded Anomaly Detection dataset: (a) Object; (b) Interaction: Applied actions shown with directional arrows; (c) Video with Physical Dynamics: Temporal sequences showing normal and abnormal states, highlighting anomalies like leaks, misalignments, and cracks. By focusing on the dynamic behaviors of complex objects, we enhance understanding of interactions and failure modes in real-world settings, where both structure and motion contribute to anomaly detection.

图1. 面向现实世界工业物体异常检测中物理动力学的视觉判别。我们展示了来自基于物理的异常检测数据集中的物体、相互作用和时间序列视频:(a) 物体；(b) 相互作用:用方向箭头表示施加的动作；(c) 带有物理动力学的视频:显示正常和异常状态的时间序列，突出显示如泄漏、错位和裂缝等异常。通过关注复杂物体的动态行为，我们增强了对现实环境中相互作用和故障模式的理解，其中结构和运动都对异常检测有贡献。

# Abstract

# 摘要

Humans detect real-world object anomalies by perceiving, interacting, and reasoning based on object-conditioned physical knowledge. The long-term goal of Industrial Anomaly Detection (IAD) is to enable machines to autonomously replicate this skill. However, current IAD algorithms are largely developed and tested on static, semantically simple datasets, which diverge from real-world scenarios where physical understanding and reasoning are essential. To bridge this gap, we introduce the Physics Anomaly Detection (Phys-AD) dataset, the first large-scale, real-world, physics-grounded video dataset for industrial anomaly detection. Collected using a real robot arm and motor, Phys-AD provides a diverse set of dynamic, semantically rich scenarios. The dataset includes more than 6400 videos across 22 real-world object categories, interacting with robot arms and motors, and exhibits 47 types of anomalies. Anomaly detection in Phys-AD requires visual reasoning, combining both physical knowledge and video content to determine object abnormality. We benchmark state-of-the-art anomaly detection methods under three settings: unsupervised AD, weakly-supervised AD, and video-understanding , highlighting their limitations in handling physics-grounded anomalies. Additionally, we introduce the Physics Anomaly Explanation (PAEval) metric, designed to assess the ability of visual-language foundation models to not only detect anomalies but also provide accurate explanations for their underlying physical causes. Our dataset and benchmark will be publicly available.

人类通过感知、相互作用和基于物体条件的物理知识来检测现实世界中的物体异常。工业异常检测(IAD)的长期目标是使机器能够自主复制这一技能。然而，当前的IAD算法主要在静态、语义简单的数据集上开发和测试，这与现实世界中需要物理理解和推理的场景相去甚远。为了弥合这一差距，我们引入了物理异常检测(Phys-AD)数据集，这是第一个用于工业异常检测的大规模、现实世界、基于物理的视频数据集。通过使用真实的机械臂和电机收集，Phys-AD提供了多样化的动态、语义丰富的场景。该数据集包括22个现实世界物体类别的6400多个视频，与机械臂和电机相互作用，并展示了47种类型的异常。在Phys-AD中进行异常检测需要视觉推理，结合物理知识和视频内容来确定物体异常。我们在三种设置下对最先进的异常检测方法进行了基准测试:无监督AD、弱监督AD和视频理解 ，突出了它们在处理基于物理的异常时的局限性。此外，我们引入了物理异常解释(PAEval)指标，旨在评估视觉语言基础模型不仅能够检测异常，还能为其潜在物理原因提供准确解释的能力。我们的数据集和基准将公开提供。

[[1]](#footnote-27)

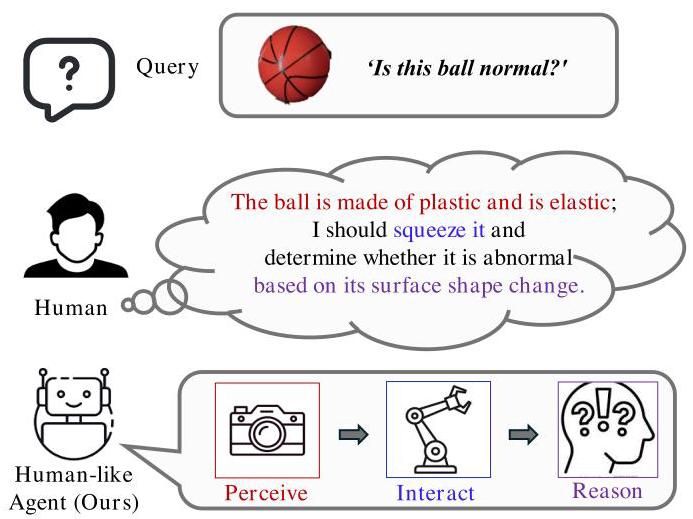


Figure 2. Human-like decision-making process for physics-grounded object anomaly detection. We illustrate the sequential approach of a human-like agent for evaluating an object’s normality. First, the agent perceives relevant physical attributes (e.g., plastic and elastic), then interacts by performing a physical action (e.g., squeezing), and finally reasons based on the vision feedback and attributes changes (e.g., surface shape change) to determine whether the object is normal or anomalous. This mirrors a human’s natural process of reasoning over physics in objects.

图2. 基于物理的物体异常检测的类人决策过程。我们展示了一个类人代理评估物体正常性的顺序方法。首先，代理感知相关的物理属性(例如，塑性和弹性)，然后通过执行物理动作(例如，挤压)进行交互，最后基于视觉反馈和属性变化(例如，表面形状变化)进行推理，以确定物体是正常还是异常。这反映了人类在物体上进行物理推理的自然过程。

# 1. Introduction

# 1. 引言

Industrial anomaly detection (IAD) is a critical subfield in computer vision and industrial automation, aiming to identify defects or irregularities in products during manufacturing. As shown in Fig. 2, the ultimate vision is to create autonomous systems that not only perceive but also interact with and reason about objects to discriminate between anomalies and normal states, integrating complex physical principles to detect anomalies in dynamic, real-world scenarios. For example, a human inspecting a water bottle for anomalies wouldn’t rely solely on visual observation; they might rotate or invert the bottle, using physical interactions and cues, such as noticing a loose cap or an irregular internal flow, to detect issues.

工业异常检测(IAD)是计算机视觉和工业自动化中的一个关键子领域，旨在识别制造过程中产品的缺陷或不规则性。如图2所示，最终愿景是创建不仅能够感知，还能与物体交互并对其进行推理以区分异常和正常状态的自主系统，整合复杂的物理原理来检测动态现实场景中的异常。例如，人类在检查水瓶的异常时，不会仅仅依赖视觉观察；他们可能会旋转或倒置瓶子，利用物理相互作用和线索，如注意到松动的盖子或不规则的内部流动，来检测问题。

Central to advancing IAD [9] is the availability of high-quality datasets that bridge the gap between academic research and industrial needs. Datasets like MVTec-AD [4], MPDD [21], and VisA [65] have played foundational roles, enabling algorithm development for image-based anomaly detection and bringing IAD to the forefront of computer vision research. While these datasets have significantly advanced single-image anomaly detection, recent datasets, e.g., MVTec-3D [7], Real3D [29], and Anomaly-ShapeNet [27], have extended IAD to 3D, aligning research more closely with the needs of complex real-world industrial settings.

推动IAD[9]发展的核心是高质量数据集的可用性，这些数据集弥合了学术研究与工业需求之间的差距。像MVTec-AD[4]、MPDD[21]和VisA[65]这样的数据集发挥了基础性作用，推动了基于图像的异常检测算法的开发，并将IAD带到了计算机视觉研究的前沿。虽然这些数据集显著推进了单图像异常检测，但最近的数据集，如MVTec-3D[7]、Real3D[29]和Anomaly-ShapeNet[27]，已将IAD扩展到3D，使研究更贴近复杂现实世界工业环境的需求。

Yet, as factories increasingly rely on robotic arms and automated systems to perform sophisticated inspections, the limitations of current IAD datasets become apparent. Existing benchmarks focus on static, semantically simple environments, overlooking the physical priors and interactive reasoning required in real-world industrial contexts. This gap highlights a growing need for datasets that not only reflect real-world physical constraints but also challenge models to reason dynamically about anomalies.

然而，随着工厂越来越多地依赖机械臂和自动化系统进行复杂的检测，当前IAD数据集的局限性变得显而易见。现有的基准测试集中在静态、语义简单的环境上，忽视了现实工业场景中所需的物理先验和交互推理。这一差距凸显了对数据集的需求，这些数据集不仅反映现实世界的物理约束，还挑战模型对异常进行动态推理。

To bridge this gap, as shown in Fig. 1, we introduce the Physics Anomaly Detection (Phys-AD) dataset, the first large-scale, physics-grounded video dataset for industrial anomaly detection. Phys-AD features over 6,400 videos of 22 categories and 49 objects interacting with robotic arms and motors, capturing 47 types of anomalies that require visual reasoning informed by physical knowledge. The short video clips in the dataset range from 60 to 240 frames in length and are filmed in real-time industrial environment, fully capturing the interaction process between robotic arm or motor and industrial objects. Additionally, to ensure our dataset meets industrial demands and matches the complexity and diversity of the real physical world, we selected industrial objects of different physical qualities, various interaction methods, and anomalies that reflect diverse physical principles and require different reasoning process. Specifically, we selected 22 object categories spanning across metals, plastics, fluid, amorphous substances and articulated objects with diverse appearances. For interaction, we use mechanical grippers, robotic arms, and motors, incorporating various interaction modes such as pressing, rotating, squeezing, and driving to handle different types of objects.

为了弥补这一差距，如图1所示，我们引入了物理异常检测(Phys-AD)数据集，这是第一个大规模、基于物理的视频数据集，用于工业异常检测。Phys-AD包含22个类别和49个物体与机械臂和电机交互的6400多个视频，捕捉了47种需要基于物理知识进行视觉推理的异常类型。数据集中的短视频片段长度从60到240帧不等，并在实时工业环境中拍摄，完整记录了机械臂或电机与工业物体之间的交互过程。此外，为了确保我们的数据集满足工业需求并匹配现实物理世界的复杂性和多样性，我们选择了具有不同物理特性的工业物体、多种交互方法以及反映不同物理原理并需要不同推理过程的异常。具体来说，我们选择了22个物体类别，涵盖金属、塑料、流体、无定形物质和具有多样外观的铰接物体。在交互方面，我们使用机械夹爪、机械臂和电机，结合按压、旋转、挤压和驱动等多种交互模式来处理不同类型的物体。

We benchmark state-of-the-art anomaly detection methods in three key configurations: unsupervised anomaly detection, weakly-supervised anomaly detection, and video-understanding based anomaly detection. Our findings reveal critical gaps in their ability to handle the complexities of physics-grounded scenarios, where anomalies often arise from dynamic, interdependent interactions. To advance the field, we also introduce the Physical Anomaly Explanation (PAEval) metric, designed to assess both detection performance and a model’s ability to explain anomalies by identifying underlying physical causes. Furthermore, our benchmark reveals the fragility of existing methods in tackling these challenging conditions, underscoring the need for approaches that better understand object dynamics and temporal coherence in anomaly detection.

我们在三种关键配置中对最先进的异常检测方法进行了基准测试:无监督异常检测、弱监督异常检测和基于视频理解的异常检测。我们的研究结果揭示了它们在处理基于物理场景复杂性方面的关键差距，这些异常通常源于动态、相互依赖的交互。为了推动该领域的发展，我们还引入了物理异常解释(PAEval)指标，旨在评估检测性能以及模型通过识别潜在物理原因来解释异常的能力。此外，我们的基准测试揭示了现有方法在处理这些挑战性条件时的脆弱性，强调了需要更好地理解物体动态和时间一致性的异常检测方法。

# Our contributions can be summarized as followings:

# 我们的贡献可以总结如下:

* We introduce a novel task of detecting physical-based industrial anomalies in real-world that involves perception, physical and visual reasoning.
* 我们引入了一项新的任务，即在现实世界中检测基于物理的工业异常，涉及感知、物理和视觉推理。
* We present Phys-AD, the first large-scale, physics-grounded video dataset specifically designed for industrial anomaly detection in real world, containing objects of different physical qualities, multiple interaction methods and various physical reasoning process.
* 我们提出了Phys-AD，这是第一个大规模、基于物理的视频数据集，专门为现实世界中的工业异常检测设计，包含具有不同物理特性的物体、多种交互方法和各种物理推理过程。
* We benchmarking the anomaly detection and reasoning performance of popular video AD methods and Visual Language Foundation Models on the Phys-AD dataset in several settings, establishing a practical and challenging benchmark to promote the development of the physics-related anomaly detection field.
* 我们在Phys-AD数据集上对流行的视频AD方法和视觉语言基础模型的异常检测和推理性能进行了基准测试，建立了一个实用且具有挑战性的基准，以促进物理相关异常检测领域的发展。

Table 1. Comparison of Phys-AD with existing industrial anomaly detection datasets. Our Phys-AD dataset is the first to consider complex objects with physical dynamics. ’Syn’, ’IR’, ’D’, and ’PC’ denote Synthetic, Infrared, Depth, and Point Cloud, respectively. #Anomaly indicates the number of anomaly types.

表1. Phys-AD与现有工业异常检测数据集的比较。我们的Phys-AD数据集首次考虑了具有物理动态的复杂物体。’Syn’、’IR’、’D’和’PC’分别表示合成、红外、深度和点云。#Anomaly表示异常类型的数量。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Year | Type | Modality | Sample Statistics | | |
|  | Class #Anomaly | Physics |
| MVTec-AD [3] | 2019 | Real | RGB | 15 | - | ✘ |
| BTAD [38] | 2021 | Real | RGB | 3 | 3 | ✘ |
| MPDD [21] | 2021 | Real | RGB | 6 | 8 | ✘ |
| VisA [66] | 2021 | Real | RGB | 12 | - | ✘ |
| MVTec LOCO-AD [5] | 2022 | Real | RGB | 5 | - | ✘ |
| MAD [64] | 2023 | Syn+Real | RGB | 20 | 3 | ✘ |
| LOCO-Annotations [62] | 2024 | Real | RGB | 5 | 5 | ✘ |
| Real-IAD [46] | 2024 | Real | RGB | 30 | 8 | ✘ |
| GDXray [37] | 2015 | Real | X-ray | 5 | 15 | ✘ |
| PVEL-AD [43] | 2023 | Real | IR | 1 | 10 | ✘ |
| MVTec3D-AD [6] | 2021 | Real | RGB-D | 10 | 3-5 | ✘ |
| Eyecandies [8] | 2022 | Syn | RGB-D | 10 | 3 | ✘ |
| Real3D-AD [30] | 2023 | Real | PC | 12 | 2 | ✘ |
| Anomaly-ShapeNet [27] | 2024 | Syn | PC | 40 | 6 | ✘ |
| Phys-AD (Ours) | 2024 | Real | RGB | 49 | 47 | ✓ |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 年份 | 类型 | 模态 | 样本统计 | | |
|  | 类别 #异常 | 物理 |
| MVTec-AD [3] | 2019 | 真实 | RGB | 15 | - | ✘ |
| BTAD [38] | 2021 | 真实 | RGB | 3 | 3 | ✘ |
| MPDD [21] | 2021 | 真实 | RGB | 6 | 8 | ✘ |
| VisA [66] | 2021 | 真实 | RGB | 12 | - | ✘ |
| MVTec LOCO-AD [5] | 2022 | 真实 | RGB | 5 | - | ✘ |
| MAD [64] | 2023 | 合成+真实 | RGB | 20 | 3 | ✘ |
| LOCO-注释 [62] | 2024 | 真实 | RGB | 5 | 5 | ✘ |
| Real-IAD [46] | 2024 | 真实 | RGB | 30 | 8 | ✘ |
| GDXray [37] | 2015 | 真实 | X射线 | 5 | 15 | ✘ |
| PVEL-AD [43] | 2023 | 真实 | 红外 | 1 | 10 | ✘ |
| MVTec3D-AD [6] | 2021 | 真实 | RGB-D | 10 | 3-5 | ✘ |
| Eyecandies [8] | 2022 | 合成 | RGB-D | 10 | 3 | ✘ |
| Real3D-AD [30] | 2023 | 真实 | 点云 | 12 | 2 | ✘ |
| Anomaly-ShapeNet [27] | 2024 | 合成 | 点云 | 40 | 6 | ✘ |
| Phys-AD (我们的) | 2024 | 真实 | RGB | 49 | 47 | ✓ |

# 2. Related work

# 2. 相关工作

Industrial anomaly detection datasets. Existing datasets primarily focus on static, semantically simple scenarios, deviate significantly from real world where physical understanding and reasoning are essential. Datasets like MVTec-AD [3], BTAD [38], MPDD [21], and VisA [66] focus on surface level image anomaly detection with one single-view RGB image, limiting their effectiveness in capturing holistic object structures. MVTec-LOCO-AD [5], focusing on structure and local information in industries, is limited by its relatively simple and constrained data content. While MVTec3D-AD [6] and Eyecandies [8] incorporate depth data, they remain static and single-view image anomaly detection, neglecting object level information. To explore object level anomaly detection, multi-view IAD datasets like MAD [64] and Real-IAD [46], point cloud IAD datasets like Real3D-AD [30] and Anomaly-ShapeNet [27], offering richer visual and geometric cues, but are still limited to static objects without dynamic interaction or reasoning. In summary, current industrial anomaly detection datasets focus on relatively simple and static anomaly detection scenarios, lacking complex physical rules, dynamic interactions, and visual reasoning requirements. Therefore, existing IAD datasets generally applied to limited industrial scenarios and there is no IAD dataset could meet the demands of detecting complex anomalies in real world which need physical priors and reasoning.

工业异常检测数据集。现有的数据集主要关注静态的、语义简单的场景，与现实中需要物理理解和推理的场景存在显著偏差。像MVTec-AD [3]、BTAD [38]、MPDD [21]和VisA [66]这样的数据集专注于单视角RGB图像的表层图像异常检测，限制了它们在捕捉整体物体结构方面的有效性。MVTec-LOCO-AD [5]专注于工业中的结构和局部信息，但其数据内容相对简单且受限。虽然MVTec3D-AD [6]和Eyecandies [8]引入了深度数据，但它们仍然是静态的单视角图像异常检测，忽略了物体层面的信息。为了探索物体层面的异常检测，像MAD [64]和Real-IAD [46]这样的多视角IAD数据集，以及像Real3D-AD [30]和Anomaly-ShapeNet [27]这样的点云IAD数据集，提供了更丰富的视觉和几何线索，但仍然局限于静态物体，缺乏动态交互或推理。总之，当前的工业异常检测数据集专注于相对简单和静态的异常检测场景，缺乏复杂的物理规则、动态交互和视觉推理需求。因此，现有的IAD数据集通常应用于有限的工业场景，没有IAD数据集能够满足检测现实世界中需要物理先验和推理的复杂异常的需求。

Video anomaly detection. Deep learning methods , 54, 61] now dominate Video Anomaly Detection (VAD), categorized into unsupervised, weakly-supervised, and fully-supervised approaches. Unsupervised methods learn normal patterns via reconstruction [16, 18, 52], prediction [32], or hybrids [33], while some methods [45, 55] train with both unlabeled normal and abnormal data. Weakly-supervised methods use video-level or glance-based annotations, and fully-supervised methods[25, 31] remain rare due to costly frame-level labeling. Visual-language models like CLIP [42] have recently been applied to enhance anomaly detection , focusing on semantic anomalies. Open-vocabulary VAD [50] and prompt-based anomaly scoring [56] leverage LLMs [53, 61], but performance relies heavily on the base models, often lacking domain-specific tuning. However, existing video anomaly detection algorithms lack the capability to handle complex industrial anomaly detection scenarios and understand physical rules. This gap highlights the need for models that can capture dynamic behaviors and physical laws in industrial environments, as addressed by Phys-AD, which targets industrial anomalies in objects of various physical properties.

视频异常检测。深度学习方法 , 54, 61] 目前在视频异常检测(VAD)中占据主导地位，分为无监督、弱监督和全监督方法。无监督方法通过重建 [16, 18, 52]、预测 [32] 或混合方法 [33] 学习正常模式，而一些方法 [45, 55] 同时使用未标记的正常和异常数据进行训练。弱监督方法 使用视频级或基于瞥见的注释，全监督方法 [25, 31] 由于帧级标注成本高而仍然罕见。像CLIP [42] 这样的视觉语言模型最近被应用于增强异常检测 ，专注于语义异常。开放词汇VAD [50] 和基于提示的异常评分 [56] 利用LLMs [53, 61]，但性能严重依赖于基础模型，通常缺乏特定领域的调优。然而，现有的视频异常检测算法缺乏处理复杂工业异常检测场景和理解物理规则的能力。这一差距凸显了需要能够捕捉工业环境中动态行为和物理规律的模型，正如Phys-AD所针对的，它专注于具有各种物理属性的物体的工业异常。

Visual reasoning. Visual reasoning is a critical task in computer vision, aiming to enable machines to interpret perceptual information like humans. Several visual reasoning tasks have been proposed to evaluate reasoning capabilities, including Visual Question Answering (VQA), 2D puzzles, and physical dynamics prediction. In VQA, agents are required to combine natural language and visual cues to answer questions . For 2D puzzles, tasks involve discovering relationships among visual elements and making inferences [22, 26, 39, 57, 58]. Physical dynamics prediction tasks require machines to perceive and reason about physical interactions [2, 12, 13, 20]. In contrast to these work, we introduce the first benchmark featuring real-world industrial objects with dynamic physical properties, focusing on distinguishing diverse dynamics through vision.

视觉推理。视觉推理是计算机视觉中的一项关键任务，旨在使机器能够像人类一样解释感知信息。已经提出了几种视觉推理任务来评估推理能力，包括视觉问答(VQA)、2D拼图和物理动力学预测。在VQA中，代理需要结合自然语言和视觉线索来回答问题 。对于2D拼图，任务涉及发现视觉元素之间的关系并进行推理 [22, 26, 39, 57, 58]。物理动力学预测任务要求机器感知和推理物理交互 [2, 12, 13, 20]。与这些工作相比，我们引入了第一个以具有动态物理属性的现实世界工业物体为特征的基准，专注于通过视觉区分不同的动态。

# 3. Dataset: Phys-AD

# 3. 数据集:Phys-AD

# 3.1. Object Preparation and Interaction Selection

# 3.1. 物体准备与交互选择

We selected 22 categories spanning across various materials, including metals, plastics, amorphous substances and articulated objects, with diverse shapes, sizes, and physical properties. For the physical properties of different objects, we correspondingly select different interaction methods like push, rotate, pull with robotics arms and motors. For instance, we use robotics arms to grab and extruding the deformable objects like basketball to determine whether there are any elastic anomalies or surface defects by the morphological change of the objects. To make our dataset more practical and challenging, we introduce 47 distinct defect types, some of which just rely on single frame content and physical rules for anomaly reasoning, while the other part needs to combine the content of the whole video and physical priors together to judge whether the object is abnormal or not. Different objects information with their corresponding interaction methods and anomalies are listed in Table 2. Fig 3 shows the interactions for understanding implicit physical laws in the Phys-AD dataset. Fig 4 provides two representative examples: one where a U Disk requires analyzing the entire video content to determine if there is an anomaly, and another where a sticky roller only requires a few frames from the video to judge whether the object is abnormal.

我们选择了22个类别，涵盖各种材料，包括金属、塑料、无定形物质和铰接物体，具有不同的形状、尺寸和物理特性。针对不同物体的物理特性，我们相应地选择了不同的交互方式，如使用机械臂和电机进行推、旋转、拉等操作。例如，我们使用机械臂抓取和挤压可变形物体(如篮球)，通过物体的形态变化来判断是否存在弹性异常或表面缺陷。为了使我们的数据集更具实用性和挑战性，我们引入了47种不同的缺陷类型，其中一些仅依赖于单帧内容和物理规则进行异常推理，而另一部分则需要结合整个视频内容和物理先验知识来判断物体是否异常。表2列出了不同物体的信息及其对应的交互方法和异常情况。图3展示了Phys-AD数据集中用于理解隐含物理定律的交互操作。图4提供了两个代表性示例:一个U盘需要分析整个视频内容以确定是否存在异常，而另一个粘性滚轮仅需要视频中的几帧即可判断物体是否异常。

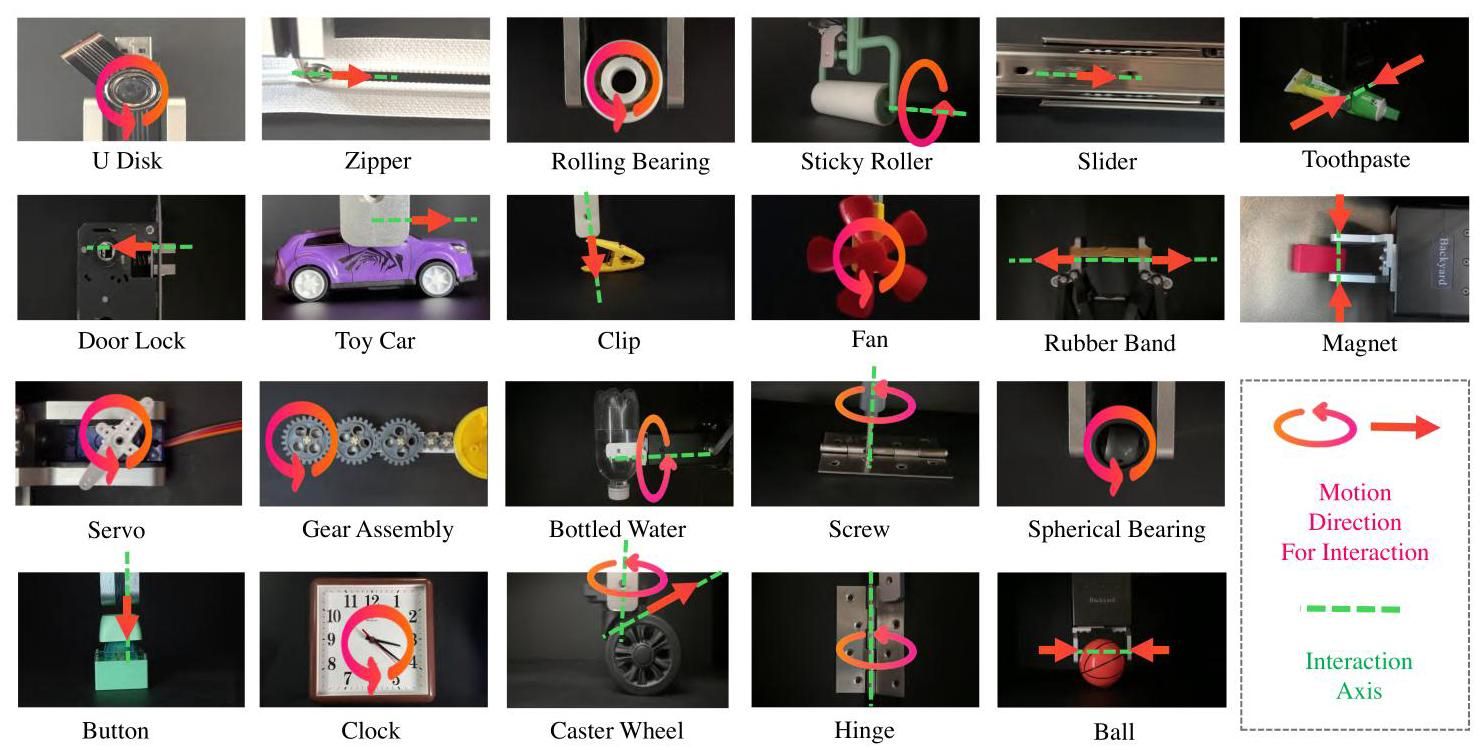


Figure 3. Interactions for understanding implicit physical laws in the Phys-AD dataset. We showcase various object interactions from the Phys-AD dataset, where different actions (indicated by motion directions) are used to explore and reason about the underlying physical properties and behaviors of each object. The colored arrows indicate the interaction directions and axes, highlighting how physical

图3. Phys-AD数据集中用于理解隐含物理定律的交互操作。我们展示了Phys-AD数据集中的各种物体交互，其中不同的动作(由运动方向表示)用于探索和推理每个物体的潜在物理特性和行为。彩色箭头表示交互方向和轴，突出显示了物理

interactions reveal the implicit physics governing each object

交互如何揭示每个物体所隐含的物理规律

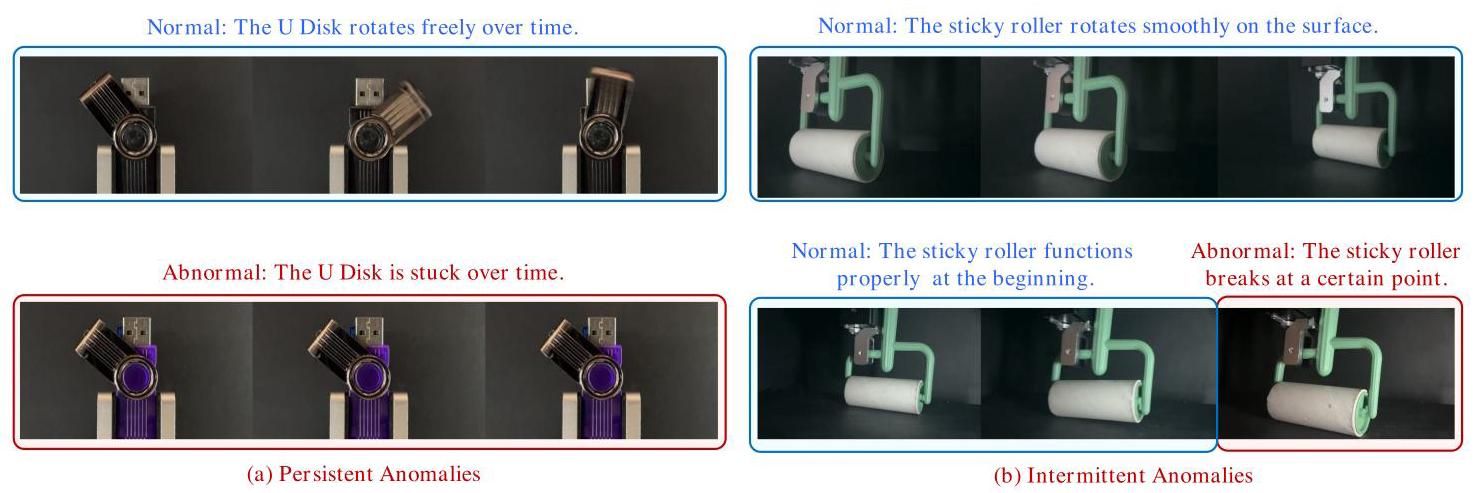


Figure 4. Categorization of anomalies based on persistence in the Phys-AD dataset. We show examples of normal and abnormal functioning in common objects, divided into two anomaly types: persistent and intermittent. (a) Persistent anomalies, such as continuous obstruction in the U Disk or permanent malfunction of the Sticky Roller, are visible throughout the operation. (b) In contrast, intermittent anomalies, like occasional jamming of the U Disk or breakage in the Sticky Roller after initial operation, only appear at specific points in time. This classification provides insight into both constant and sporadic failures in object interactions.

图4. 基于持续性的Phys-AD数据集异常分类。我们展示了常见物体的正常和异常功能示例，分为两种异常类型:持续性和间歇性。(a) 持续性异常，如U盘的持续阻塞或粘性滚轮的永久故障，在整个操作过程中可见。(b) 相比之下，间歇性异常，如U盘的偶尔卡住或粘性滚轮在初始操作后的断裂，仅在特定时间点出现。这种分类为物体交互中的持续性和偶发性故障提供了洞察。

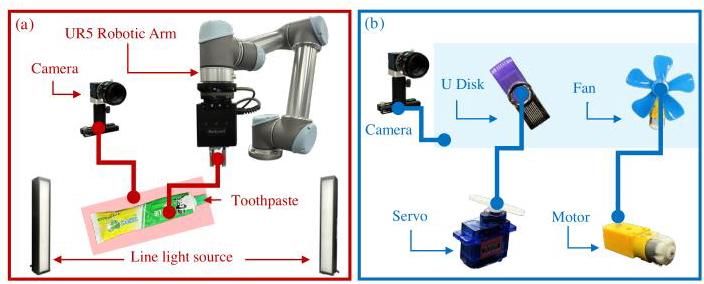


Figure 5. Data collection pipeline for the Phys-AD dataset. (a) Manipulation of a toothpaste tube using a UR5 robotic arm. (b) Manipulation of a U Disk and fan via servo and motor.

图5. Phys-AD数据集的数据收集流程。(a) 使用UR5机械臂操作牙膏管。(b) 通过伺服和电机操作U盘和风扇。

Table 2. Interaction methods and anomalies of the object categories in our Phys-AD dataset. Motor. means one object driven by motor. Multi. means one object with multiple anomalies.

表2. 我们Phys-AD数据集中物体类别的交互方法和异常情况。Motor. 表示由电机驱动的物体。Multi. 表示具有多个异常的物体。

|  |  |  |
| --- | --- | --- |
| Category | Interaction | Anomaly Types |
| Car | Drag, Slide | Different wheels stuck |
| Fan | Motor., Rotate | Stuck, Uneven Rotation, Vibration |
| Rolling Bearing | Motor., Rotate | Lack of friction |
| Spherical Bearing | Grab, Rotate | Internal block |
| Servo | Grab, Motor., Rotate | Angle restricted, Vibration, No calibration |
| Clip | Press | Unable to press down, Unable to rebound |
| U Disk | Grab, Motor., Rotate | Cover jam |
| Hinge | Grab, Rotate | Angle restricted, Shaft off |
| Sticky Roller | Grab, Pull | Detach, Unable to rotate |
| Caster Wheel | Slide | Axle axis stuck, Swivel axis stuck |
| Screw | Press, Rotate | Loosen, Unable to insert |
| Lock | Motor., Rotate | Latch jam, Loose |
| Gear | Motor., Rotate | Stuck, Not meshed, Multi. |
| Clock | Motor., Rotate | Pointer stops, Vibration |
| Slide | Grab, Slide | Detach, Shedding, Jam |
| Zipper | Grab, Close | Stuck, Unable to close |
| Button | Press | No light, Unable to press down, Unable to rebound, Multi. |
| Liquid | Grab, Shake | Water out, Foreign body |
| Rubber Band | Stretch | Crack |
| Ball | Pinch, Press | Insufficient gas, Leakage |
| Magnet | Grab, Press, Move | Degaussing, Shell detached |
| Toothpaste | Pinch, Press | Leakage |

|  |  |  |
| --- | --- | --- |
| 类别 | 交互 | 异常类型 |
| 汽车 | 拖动，滑动 | 不同车轮卡住 |
| 风扇 | 电机，旋转 | 卡住，旋转不均匀，振动 |
| 滚动轴承 | 电机，旋转 | 缺乏摩擦 |
| 球面轴承 | 抓取，旋转 | 内部阻塞 |
| 伺服 | 抓取，电机，旋转 | 角度受限，振动，未校准 |
| 夹子 | 按压 | 无法按下，无法回弹 |
| U盘 | 抓取，电机，旋转 | 盖子卡住 |
| 铰链 | 抓取，旋转 | 角度受限，轴脱落 |
| 粘性滚轮 | 抓取，拉动 | 分离，无法旋转 |
| 脚轮 | 滑动 | 轴卡住，旋转轴卡住 |
| 螺丝 | 按压，旋转 | 松动，无法插入 |
| 锁 | 电机，旋转 | 锁扣卡住，松动 |
| 齿轮 | 电机，旋转 | 卡住，未啮合，多重 |
| 时钟 | 电机，旋转 | 指针停止，振动 |
| 滑动 | 抓取，滑动 | 分离，脱落，卡住 |
| 拉链 | 抓取，关闭 | 卡住，无法关闭 |
| 按钮 | 按压 | 无光，无法按下，无法回弹，多重 |
| 液体 | 抓取，摇晃 | 出水，异物 |
| 橡皮筋 | 拉伸 | 裂缝 |
| 球 | 捏，按压 | 气体不足，泄漏 |
| 磁铁 | 抓取，按压，移动 | 消磁，外壳脱落 |
| 牙膏 | 捏，按压 | 泄漏 |

# 3.2. Data Collection and Processing

# 3.2. 数据收集与处理

Data collection pipeline. Most of the data collection for the Phys-AD dataset is driven by manipulation-guided video sequences, captured using a UR5 robotic arm equipped with an RGB camera (see Fig. 5a). In order to reproduce real industrial scenes, we also include adjustable light sources in our data capture process. The RGB Camera, with a resolution and a frame rate of , providing high-quality video sequences. Some kinds of the objects like U Disk that are not suitable for the manipulation of robotics arms, are driven by the motor or servo (Fig. 5b). After the data collection, we used video editing software to crop out irrelevant frames and retain the complete interaction process.

数据收集流程。Phys-AD数据集的大部分数据收集是通过操作引导的视频序列进行的，这些视频序列使用配备RGB摄像头的UR5机械臂捕获(见图5a)。为了再现真实的工业场景，我们还在数据捕获过程中加入了可调节的光源。RGB摄像头具有 分辨率和 帧率，提供高质量的视频序列。一些不适合机械臂操作的对象，如U盘，由电机或伺服驱动(图5b)。数据收集完成后，我们使用视频编辑软件裁剪掉无关的帧，并保留完整的交互过程。

Table 3. Phys-AD dataset statistics. We denote the total number of frames and videos for each category as #Images and #Videos. Note that the Train split does not contain anomaly data.

表3. Phys-AD数据集统计。我们将每个类别的总帧数和视频数表示为#Images和#Videos。请注意，训练集不包含异常数据。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Train | Train | Test | Test | Test #Videos Nor. | Test #Videos Ab. | Types | Types |
| Car | 18,000 | 300 | 36,000 | 600 | 150 | 450 | 5 | 3 |
| Fan | 32,400 | 180 | 64,800 | 360 | 90 | 270 | 3 | 3 |
| Rolling Bearing | 3,600 | 60 | 3,600 | 60 | 30 | 30 | 1 | 1 |
| Spherical Bearing | 3,600 | 60 | 3,600 | 60 | 30 | 30 | 1 | 1 |
| Servo | 14,400 | 120 | 28,800 | 240 | 60 | 180 | 1 | 3 |
| Clip | 28,800 | 240 | 43,200 | 360 | 120 | 240 | 4 | 2 |
| U Disk | 14,400 | 240 | 14,400 | 240 | 120 | 120 | 4 | 1 |
| Hinge | 3,600 | 30 | 7,200 | 60 | 15 | 45 | 1 | 2 |
| Sticky Roller | 5,400 | 30 | 8,100 | 45 | 15 | 30 | 1 | 2 |
| Caster Wheel | 5,400 | 30 | 10,800 | 60 | 15 | 45 | 1 | 2 |
| Screw | 5,400 | 30 | 8,100 | 45 | 15 | 30 | 1 | 2 |
| Lock | 7,200 | 120 | 10,800 | 180 | 60 | 120 | 1 | 2 |
| Gear | 21,600 | 180 | 54,000 | 450 | 90 | 360 | 3 | 4 |
| Clock | 27,000 | 150 | 40,320 | 224 | 73 | 151 | 5 | 2 |
| Slide | 7,200 | 60 | 18,000 | 150 | 30 | 120 | 1 | 3 |
| Zipper | 14,400 | 120 | 21,600 | 180 | 60 | 120 | 2 | 2 |
| Button | 21,600 | 120 | 54,000 | 300 | 60 | 240 | 4 | 4 |
| Liquid | 5,400 | 30 | 8,100 | 45 | 15 | 30 | 1 | 2 |
| Rubber Band | 10,800 | 60 | 10,800 | 60 | 30 | 30 | 1 | 1 |
| Ball | 21,600 | 90 | 32,400 | 135 | 45 | 30 | 3 | 2 |
| Magnet | 10,800 | 60 | 16,200 | 90 | 30 | 60 | 2 | 2 |
| Toothpaste | 16,200 | 90 | 16,200 | 90 | 45 | 45 | 3 | 1 |
| Total | 298,800 | 2400 | 511,020 | 4,034 | 1,198 | 2,836 | 49 | 47 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类别 | 训练 | 训练 | 测试 | 测试 | 测试 #视频数量 | 测试 #视频异常 | 类型 | 类型 |
| 汽车 | 18,000 | 300 | 36,000 | 600 | 150 | 450 | 5 | 3 |
| 风扇 | 32,400 | 180 | 64,800 | 360 | 90 | 270 | 3 | 3 |
| 滚动轴承 | 3,600 | 60 | 3,600 | 60 | 30 | 30 | 1 | 1 |
| 球面轴承 | 3,600 | 60 | 3,600 | 60 | 30 | 30 | 1 | 1 |
| 伺服 | 14,400 | 120 | 28,800 | 240 | 60 | 180 | 1 | 3 |
| 夹子 | 28,800 | 240 | 43,200 | 360 | 120 | 240 | 4 | 2 |
| U盘 | 14,400 | 240 | 14,400 | 240 | 120 | 120 | 4 | 1 |
| 铰链 | 3,600 | 30 | 7,200 | 60 | 15 | 45 | 1 | 2 |
| 粘性滚轮 | 5,400 | 30 | 8,100 | 45 | 15 | 30 | 1 | 2 |
| 脚轮 | 5,400 | 30 | 10,800 | 60 | 15 | 45 | 1 | 2 |
| 螺丝 | 5,400 | 30 | 8,100 | 45 | 15 | 30 | 1 | 2 |
| 锁 | 7,200 | 120 | 10,800 | 180 | 60 | 120 | 1 | 2 |
| 齿轮 | 21,600 | 180 | 54,000 | 450 | 90 | 360 | 3 | 4 |
| 时钟 | 27,000 | 150 | 40,320 | 224 | 73 | 151 | 5 | 2 |
| 滑轨 | 7,200 | 60 | 18,000 | 150 | 30 | 120 | 1 | 3 |
| 拉链 | 14,400 | 120 | 21,600 | 180 | 60 | 120 | 2 | 2 |
| 按钮 | 21,600 | 120 | 54,000 | 300 | 60 | 240 | 4 | 4 |
| 液体 | 5,400 | 30 | 8,100 | 45 | 15 | 30 | 1 | 2 |
| 橡皮筋 | 10,800 | 60 | 10,800 | 60 | 30 | 30 | 1 | 1 |
| 球 | 21,600 | 90 | 32,400 | 135 | 45 | 30 | 3 | 2 |
| 磁铁 | 10,800 | 60 | 16,200 | 90 | 30 | 60 | 2 | 2 |
| 牙膏 | 16,200 | 90 | 16,200 | 90 | 45 | 45 | 3 | 1 |
| 总计 | 298,800 | 2400 | 511,020 | 4,034 | 1,198 | 2,836 | 49 | 47 |

# 3.3. Data Statistics

# 3.3. 数据统计

Dataset sample distribution. Table 3 provides a detailed breakdown of the Phys-AD dataset, which includes information on category, the number of training videos and frames, the number of testing videos and frames, the distribution of normal and abnormal samples in the test set, and the number of object and defect types. The length of the video clips in the dataset ranges from 60 to 240 frames, ensuring to fully capture the interaction process in a short time. The dataset contains 2400 training videos and 4034 test videos spreading across 22 categories, 49 object types, and 47 defect types. The frames are extracted from the videos at a rate of 60 FPS. The test set includes 1,198 normal samples and 2,836 abnormal samples. Overall, the Phys-AD dataset covers a wide range of categories and contains a large scale of data, which helps to train more robust anomaly detection models and provides a reasonable evaluation setting.

数据集样本分布。表3详细列出了Phys-AD数据集的分类信息，包括类别、训练视频和帧的数量、测试视频和帧的数量、测试集中正常和异常样本的分布，以及物体和缺陷类型的数量。数据集中的视频片段长度从60帧到240帧不等，确保在短时间内充分捕捉交互过程。数据集包含2400个训练视频和4034个测试视频，涵盖22个类别、49种物体类型和47种缺陷类型。帧从视频中以60 FPS的速率提取。测试集包括1198个正常样本和2836个异常样本。总体而言，Phys-AD数据集覆盖了广泛的类别，并包含大规模的数据，有助于训练更稳健的异常检测模型，并提供合理的评估设置。

Interaction and defect types across categories. Table 2 details the defect types and corresponding interaction methods for each category in the Phys-AD dataset. Covering 22 categories and including 47 defect types, the dataset features various interactions (e.g., Rotate, Grab). Importantly, interaction methods or defect types with identical names can differ across categories-for example, "rotation" for a fan describes the motor-driven movement of its blades, while for a hinge it refers to the turning of its pages around the shaft under robotic control. Furthermore, to detect anomalies with complex physical properties in a human-like manner, we sometimes combine multiple interactions. For instance, when inspecting a screw, we first press it into the hole and then use a robotic arm to rotate it further. Overall, Phys-AD provides a diverse, physics-grounded anomaly detection framework that fosters the development of advanced methods for tackling real-world challenges.

跨类别的交互和缺陷类型。表2详细列出了Phys-AD数据集中每个类别的缺陷类型和相应的交互方法。涵盖22个类别和47种缺陷类型，数据集包含各种交互(例如，旋转、抓取)。重要的是，不同类别中名称相同的交互方法或缺陷类型可能有所不同——例如，风扇的“旋转”描述的是其叶片由电机驱动的运动，而铰链的“旋转”则指的是在机器人控制下其页面围绕轴的转动。此外，为了以类似人类的方式检测具有复杂物理特性的异常，我们有时会结合多种交互。例如，在检查螺丝时，我们首先将其压入孔中，然后使用机械臂进一步旋转它。总体而言，Phys-AD提供了一个多样化、基于物理的异常检测框架，促进了应对现实世界挑战的先进方法的发展。

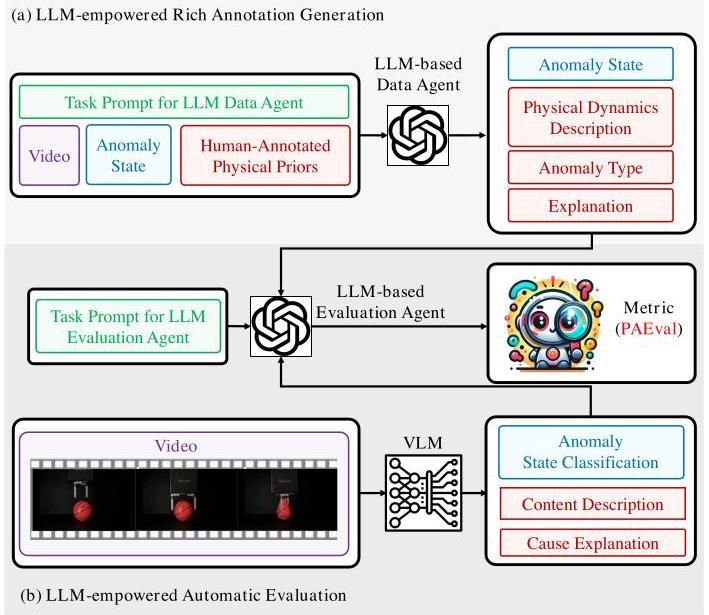


Figure 6. PhysAD-Agent: A Large Language Model (LLM)- powered system for physics anomaly detection label augmentation and automatic evaluation. This agent framework consists of two main components: (a) Rich Annotation Generation, where an LLM-based data agent generates detailed anomaly annotations based on video, anomaly state, prompt, and human-annotated physical priors, and (b) Automatic Evaluation, where an LLM-based evaluation agent assesses model predictions to calculate the Physics Anomaly Explanation (PAEval) metric.

图6. PhysAD-Agent:一个基于大语言模型(LLM)的物理异常检测标签增强和自动评估系统。该代理框架由两个主要组件组成:(a)丰富注释生成，其中基于LLM的数据代理根据视频、异常状态、提示和人工标注的物理先验生成详细的异常注释；(b)自动评估，其中基于LLM的评估代理评估模型预测以计算物理异常解释(PAEval)指标。

Labels. For the test set videos, we provide two types of labels. First, there are the common video-level labels: all anomalous videos in the test set are labeled as 1, while normal videos are labeled as 0 . For evaluating visual-language models, we also provide text labels. Specifically, for each type of anomaly in the test set, we manually design a textual description label. This label includes both a description of the video content and a physical explanation of the reason for the anomaly. To ensure diversity in our text labels, we also use ChatGPT-4o for text augmentation. For further details on the labeling process, please refer to Fig 6a.

标签。对于测试集视频，我们提供了两种类型的标签。首先，常见的视频级标签:测试集中的所有异常视频都标记为1，而正常视频标记为0。为了评估视觉语言模型，我们还提供了文本标签。具体来说，对于测试集中的每种异常类型，我们手动设计了一个文本描述标签。该标签包括对视频内容的描述和对异常原因的物理解释。为了确保文本标签的多样性，我们还使用ChatGPT-4o进行文本增强。有关标注过程的更多详细信息，请参见图6a。

# 4. Phys-AD Benchmark

# 4. Phys-AD基准

# 4.1. Problem Definition and Challenges

# 4.1. 问题定义与挑战

We establish unsupervised and weakly supervised AD settings for Phys-AD, using unsupervised as the default in our problem definition.

我们为Phys-AD建立了无监督和弱监督的异常检测设置，并在问题定义中默认使用无监督设置。

We formulate our Phys-AD setting as two steps.

我们将Phys-AD设置分为两个步骤。

Step 1: Rules deduction Given a set of training objects from category , we first use robotics arms and motors to interact with in corresponding methods , and we get the interaction process in video format as . Then, we feed the video sequence and the corresponding category’s physical prior together to the deduction function to obtain the normal rules . After rules deduction for all the categories, we get the rules bank . Step 1 can be formulated as below:

步骤1:规则推导 给定来自类别 的一组训练对象 ，我们首先使用机械臂和电机以相应的方法 与 进行交互，并将交互过程以视频格式 记录下来。然后，我们将视频序列 和相应类别的物理先验 一起输入到推导函数 中，以获得正常规则 。在对所有类别进行规则推导后，我们得到规则库 。步骤1可以表述如下:

represents interaction between objects and robotics arms. is the total number of categories.

表示物体与机械臂之间的交互。 是类别总数。

Step 2: Anomaly reasoning During test time, we first use robotics arms and motors to interact with from the test set in corresponding methods , and we get the interaction process in video format as . Then, we feed the video sequence and rules from step 1 to the reasoner to predict the object’s anomaly score . Step 2 can be formulated as:

步骤2:异常推理 在测试阶段，我们首先使用机械臂和电机以相应的方法 与测试集中的 进行交互，并将交互过程以视频格式 记录下来。然后，我们将视频序列 和步骤1中的规则 输入到推理器 中，以预测对象的异常分数 。步骤2可以表示为:

Challenges. Key challenges include combining video information with objects’ physical priors to deduct normal rules, capturing long-term temporal dependencies and fine-grained frame-level information, ensuring robust reasoning for anomaly prediction, and generalizing across diverse anomaly patterns. Addressing these challenges is essential for advancing anomaly detection in real-world environment.

挑战。关键挑战包括将视频信息与物体的物理先验知识结合以推导正常规则，捕捉长期时间依赖性和细粒度的帧级信息，确保异常预测的鲁棒推理，以及在不同异常模式之间进行泛化。解决这些挑战对于推进现实环境中的异常检测至关重要。

# 4.2. Evaluation Settings

# 4.2. 评估设置

Unsupervised AD. Unsupervised AD is the most common IAD setting for existing IAD datasets and algorithms. Under unsupervised AD setting, the training set contains only normal video data, and the algorithm needs to capture the normal distribution of the data from the training set. In the test set, both normal and abnormal data are included, and the algorithm must distinguish between normal and abnormal data based on the distribution learned during training.

无监督异常检测(Unsupervised AD)。无监督异常检测是现有IAD数据集和算法中最常见的IAD设置。在无监督异常检测设置下，训练集仅包含正常视频数据，算法需要从训练集中捕捉数据的正常分布。在测试集中，既包含正常数据也包含异常数据，算法必须根据训练期间学到的分布来区分正常和异常数据。

Weakly-supervised AD. Our dataset is in the form of videos for complex anomaly detection in industrial scenarios. In the video context, it is inevitable that we need to discuss weakly supervised anomaly detection. Under weakly-supervised Phys-AD setting, 24̃ video-level labeled anomaly samples are sampled from all possible anomaly classes in the test set in our Phys-AD dataset. These sampled anomalies are then removed from the test set. It’s worth noting that we only provide video level label in our test set. This is because most anomalies in our test set require reasoning based on the entire video content and physical knowledge of the objects.

弱监督异常检测(Weakly-supervised AD)。我们的数据集以视频形式呈现，用于工业场景中的复杂异常检测。在视频背景下，不可避免地需要讨论弱监督异常检测。在弱监督Phys-AD设置下，我们从Phys-AD数据集的测试集中所有可能的异常类别中采样24̃个视频级标记的异常样本。然后，这些采样的异常样本从测试集中移除。值得注意的是，我们仅在测试集中提供视频级标签。这是因为我们测试集中的大多数异常需要基于整个视频内容和物体的物理知识进行推理。

Video-Understanding AD. Video-Understanding models are another potential solution to our Phys-AD setting. Similar to unsupervised setting, we only provide normal videos during training. The visual language models need to provide explicit normal rules in language format during training. In the test phase, these video-understanding AD methods have to truly understand the videos and predict the anomalies based on the language rules withdrawn from the training phase.

视频理解异常检测(Video-Understanding AD)。视频理解模型是我们Phys-AD设置的另一种潜在解决方案。与无监督设置类似，我们在训练期间仅提供正常视频。视觉语言模型需要在训练期间以语言格式提供明确的正常规则。在测试阶段，这些视频理解异常检测方法必须真正理解视频，并根据训练阶段提取的语言规则预测异常。

Table 4. Video-level AUROC ( ) result of 22 categories on Phys-AD dataset. We include Unsupervised, Weakly-supervised and Video-understanding methods. The best per-category result for each class of methods is highlighted in bold. "ZS ImgB", "V-ChatGPT, "V-LLaMA’,’V-LLaVA’ denote ZS ImageBind, Video-ChatGPT, Video-LLaMA and Video-LLaVA.

表4. Phys-AD数据集上22个类别的视频级AUROC( )结果。我们包括无监督、弱监督和视频理解方法。每类方法的最佳结果以粗体突出显示。"ZS ImgB"、"V-ChatGPT"、"V-LLaMA"、"V-LLaVA"分别表示ZS ImageBind、Video-ChatGPT、Video-LLaMA和Video-LLaVA。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category. | Unsupervised | | | | | Weakly-supervised | | | Video-understanding | | | | | |
| MPN [34] | MemAE [16] | MNAD.p [40] | MNAD.r [40] | SVM [44] | VADClip [51] | S3R [48] | MGFN [10] | LAVAD [56] | ZS Clip [42] | ZS ImgB [15] | V-ChatGPT [36] | V-LLaMA [59] | V-LLaVA [28] |
| Car | 0.229 | 0.523 | 0.492 | 0.944 | 0.587 | 0.581 | 0.606 | 0.571 | 0.557 | 0.500 | 0.500 | 0.500 | 0.678 | 0.522 |
| Fan | 0.811 | 0.371 | 0.810 | 0.542 | 0.500 | 0.624 | 0.640 | 0.542 | 0.510 | 0.500 | 0.500 | 0.549 | 0.592 | 0.611 |
| Rolling Bearing | 0.353 | 0.044 | 0.352 | 0.800 | 0.933 | 0.589 | 0.601 | 0.680 | 0.532 | 0.500 | 0.500 | 0.300 | 0.500 | 0.500 |
| Spherical Bearing | 0.113 | 0.092 | 0.962 | 0.813 | 0.650 | 0.500 | 0.682 | 0.583 | 0.435 | 0.500 | 0.500 | 0.450 | 0.550 | 0.500 |
| Servo | 0.364 | 0.445 | 0.975 | 0.878 | 0.500 | 0.518 | 0.592 | 0.556 | 0.502 | 0.500 | 0.500 | 0.506 | 0.683 | 0.464 |
| Clip | 0.535 | 0.443 | 0.630 | 0.333 | 0.500 | 0.412 | 0.561 | 0.563 | 0.516 | 0.500 | 0.500 | 0.669 | 0.556 | 0.458 |
| U Disk | 0.240 | 0.617 | 0.609 | 0.940 | 0.500 | 0.530 | 0.549 | 0.570 | 0.513 | 0.500 | 0.500 | 0.565 | 0.575 | 0.500 |
| Hinge | 0.769 | 0.870 | 0.705 | 0.895 | 0.500 | 0.737 | 0.561 | 0.550 | 0.564 | 0.500 | 0.500 | 0.500 | 0.692 | 0.500 |
| Sticky Roller | 0.967 | 0.967 | 0.451 | 0.936 | 0.500 | 0.542 | 0.835 | 0.669 | 0.266 | 0.500 | 0.500 | 0.450 | 0.544 | 0.467 |
| Caster Wheel | 0.364 | 0.523 | 0.271 | 0.508 | 0.500 | 0.587 | 0.676 | 0.676 | 0.615 | 0.500 | 0.500 | 0.444 | 0.642 | 0.500 |
| Screw | 0.522 | 0.567 | 0.680 | 0.547 | 0.500 | 0.500 | 0.657 | 0.541 | 0.688 | 0.500 | 0.500 | 0.472 | 0.256 | 0.550 |
| Lock | 0.563 | 0.597 | 0.430 | 0.641 | 0.733 | 0.452 | 0.523 | 0.626 | 0.341 | 0.500 | 0.500 | 0.279 | 0.494 | 0.500 |
| Gear | 0.529 | 0.510 | 0.652 | 0.694 | 0.500 | 0.519 | 0.580 | 0.544 | 0.603 | 0.500 | 0.500 | 0.544 | 0.603 | 0.517 |
| Clock | 0.340 | 0.542 | 0.395 | 0.587 | 0.500 | 0.500 | 0.549 | 0.509 | 0.500 | 0.500 | 0.500 | 0.501 | 0.360 | 0.500 |
| Slide | 0.517 | 0.962 | 0.917 | 0.784 | 0.500 | 0.531 | 0.611 | 0.568 | 0.425 | 0.500 | 0.500 | 0.562 | 0.567 | 0.179 |
| Zipper | 0.815 | 0.592 | 0.829 | 0.421 | 0.500 | 0.504 | 0.636 | 0.633 | 0.535 | 0.500 | 0.500 | 0.547 | 0.489 | 0.500 |
| Button | 0.853 | 0.365 | 0.660 | 0.568 | 0.500 | 0.627 | 0.515 | 0.566 | 0.439 | 0.500 | 0.500 | 0.515 | 0.517 | 0.360 |
| Liquid | 0.184 | 0.700 | 0.671 | 0.831 | 0.500 | 0.542 | 0.595 | 0.793 | 0.504 | 0.500 | 0.500 | 0.410 | 0.278 | 0.217 |
| Rubber Band | 0.374 | 0.368 | 0.366 | 0.307 | 0.567 | 0.482 | 0.623 | 0.604 | 0.511 | 0.500 | 0.500 | 0.517 | 0.450 | 0.450 |
| Ball | 0.543 | 0.383 | 0.728 | 0.687 | 0.500 | 0.500 | 0.667 | 0.567 | 0.603 | 0.500 | 0.500 | 0.562 | 0.636 | 0.533 |
| Magnet | 0.671 | 0.464 | 0.691 | 0.438 | 0.500 | 0.500 | 0.548 | 0.719 | 0.502 | 0.500 | 0.500 | 0.683 | 0.300 | 0.400 |
| Toothpaste | 0.587 | 0.889 | 0.520 | 0.631 | 0.500 | 0.500 | 0.686 | 0.711 | 0.562 | 0.500 | 0.500 | 0.376 | 0.550 | 0.467 |
| Average | 0.511 | 0.538 | 0.627 | 0.669 | 0.544 | 0.535 | 0.613 | 0.606 | 0.510 | 0.500 | 0.500 | 0.496 | 0.523 | 0.463 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类别。 | 无监督 | | | | | 弱监督 | | | 视频理解 | | | | | |
| MPN [34] | MemAE [16] | MNAD.p [40] | MNAD.r [40] | SVM [44] | VADClip [51] | S3R [48] | MGFN [10] | LAVAD [56] | ZS Clip [42] | ZS ImgB [15] | V-ChatGPT [36] | V-LLaMA [59] | V-LLaVA [28] |
| 汽车 | 0.229 | 0.523 | 0.492 | 0.944 | 0.587 | 0.581 | 0.606 | 0.571 | 0.557 | 0.500 | 0.500 | 0.500 | 0.678 | 0.522 |
| 风扇 | 0.811 | 0.371 | 0.810 | 0.542 | 0.500 | 0.624 | 0.640 | 0.542 | 0.510 | 0.500 | 0.500 | 0.549 | 0.592 | 0.611 |
| 滚动轴承 | 0.353 | 0.044 | 0.352 | 0.800 | 0.933 | 0.589 | 0.601 | 0.680 | 0.532 | 0.500 | 0.500 | 0.300 | 0.500 | 0.500 |
| 球面轴承 | 0.113 | 0.092 | 0.962 | 0.813 | 0.650 | 0.500 | 0.682 | 0.583 | 0.435 | 0.500 | 0.500 | 0.450 | 0.550 | 0.500 |
| 伺服 | 0.364 | 0.445 | 0.975 | 0.878 | 0.500 | 0.518 | 0.592 | 0.556 | 0.502 | 0.500 | 0.500 | 0.506 | 0.683 | 0.464 |
| 夹子 | 0.535 | 0.443 | 0.630 | 0.333 | 0.500 | 0.412 | 0.561 | 0.563 | 0.516 | 0.500 | 0.500 | 0.669 | 0.556 | 0.458 |
| U盘 | 0.240 | 0.617 | 0.609 | 0.940 | 0.500 | 0.530 | 0.549 | 0.570 | 0.513 | 0.500 | 0.500 | 0.565 | 0.575 | 0.500 |
| 铰链 | 0.769 | 0.870 | 0.705 | 0.895 | 0.500 | 0.737 | 0.561 | 0.550 | 0.564 | 0.500 | 0.500 | 0.500 | 0.692 | 0.500 |
| 粘性滚轮 | 0.967 | 0.967 | 0.451 | 0.936 | 0.500 | 0.542 | 0.835 | 0.669 | 0.266 | 0.500 | 0.500 | 0.450 | 0.544 | 0.467 |
| 脚轮 | 0.364 | 0.523 | 0.271 | 0.508 | 0.500 | 0.587 | 0.676 | 0.676 | 0.615 | 0.500 | 0.500 | 0.444 | 0.642 | 0.500 |
| 螺丝 | 0.522 | 0.567 | 0.680 | 0.547 | 0.500 | 0.500 | 0.657 | 0.541 | 0.688 | 0.500 | 0.500 | 0.472 | 0.256 | 0.550 |
| 锁 | 0.563 | 0.597 | 0.430 | 0.641 | 0.733 | 0.452 | 0.523 | 0.626 | 0.341 | 0.500 | 0.500 | 0.279 | 0.494 | 0.500 |
| 齿轮 | 0.529 | 0.510 | 0.652 | 0.694 | 0.500 | 0.519 | 0.580 | 0.544 | 0.603 | 0.500 | 0.500 | 0.544 | 0.603 | 0.517 |
| 时钟 | 0.340 | 0.542 | 0.395 | 0.587 | 0.500 | 0.500 | 0.549 | 0.509 | 0.500 | 0.500 | 0.500 | 0.501 | 0.360 | 0.500 |
| 滑轨 | 0.517 | 0.962 | 0.917 | 0.784 | 0.500 | 0.531 | 0.611 | 0.568 | 0.425 | 0.500 | 0.500 | 0.562 | 0.567 | 0.179 |
| 拉链 | 0.815 | 0.592 | 0.829 | 0.421 | 0.500 | 0.504 | 0.636 | 0.633 | 0.535 | 0.500 | 0.500 | 0.547 | 0.489 | 0.500 |
| 按钮 | 0.853 | 0.365 | 0.660 | 0.568 | 0.500 | 0.627 | 0.515 | 0.566 | 0.439 | 0.500 | 0.500 | 0.515 | 0.517 | 0.360 |
| 液体 | 0.184 | 0.700 | 0.671 | 0.831 | 0.500 | 0.542 | 0.595 | 0.793 | 0.504 | 0.500 | 0.500 | 0.410 | 0.278 | 0.217 |
| 橡皮筋 | 0.374 | 0.368 | 0.366 | 0.307 | 0.567 | 0.482 | 0.623 | 0.604 | 0.511 | 0.500 | 0.500 | 0.517 | 0.450 | 0.450 |
| 球 | 0.543 | 0.383 | 0.728 | 0.687 | 0.500 | 0.500 | 0.667 | 0.567 | 0.603 | 0.500 | 0.500 | 0.562 | 0.636 | 0.533 |
| 磁铁 | 0.671 | 0.464 | 0.691 | 0.438 | 0.500 | 0.500 | 0.548 | 0.719 | 0.502 | 0.500 | 0.500 | 0.683 | 0.300 | 0.400 |
| 牙膏 | 0.587 | 0.889 | 0.520 | 0.631 | 0.500 | 0.500 | 0.686 | 0.711 | 0.562 | 0.500 | 0.500 | 0.376 | 0.550 | 0.467 |
| 平均 | 0.511 | 0.538 | 0.627 | 0.669 | 0.544 | 0.535 | 0.613 | 0.606 | 0.510 | 0.500 | 0.500 | 0.496 | 0.523 | 0.463 |

# 4.3. Evaluation Metrics

# 4.3. 评估指标

Physics anomaly accuracy metrics. We use the Area Under the Receiver Operating Characteristic Curve (AUROC) to evaluate video-level anomaly detection performance. We also report the average precision (AP), i.e. the area under the video-level precision-recall curve and the acc (ACCURACY), following previous works [47, 49].

物理异常准确率指标。我们使用接收者操作特征曲线下面积(AUROC)来评估视频级异常检测性能。我们还报告了平均精度(AP)，即视频级精度-召回曲线下的面积，以及准确率(ACCURACY)，遵循之前的工作[47, 49]。

Physics Anomaly Explanation (PAEval) metric. Directly utilizing existing VLMs [59][36] to understand the videos in our dataset and detect the anomalies is a potential solution to our challenge. The key point is: Can existing VLMs truly understand physical rules and reason in a right way? Specifically designed for video-understanding methods, we introduce a new metric named Physics Anomaly Explanation (PAEval) metric. To be more specific, PAEval evaluate the anomaly detection performance of algorithms based on VLMs from three different perspectives: classification, description, and explanation. Classification refers to the traditional anomaly detection metrics like AUROC, etc. Inspired by the works , PAEval also introduces two additional evaluation metrics: description and physical explanation. Description refers to the model’s ability to correctly describe the content of the video, used to assess whether the model has the capability to describe physical phenomena. Explanation refers to the model’s ability to correctly explain the physical causes of anomalies in the video, used to evaluate the VLM’s reasoning ability. To provide labels for description and explanation metrics, we manually labeled each type of defects from all categories and performed data augmentation by ChatGPT to ensure the robustness of the detection. The whole pipeline of PAEval metric is depicted in Fig 6.

物理异常解释(PAEval)指标。直接利用现有的视觉语言模型(VLMs)[59][36]来理解我们数据集中的视频并检测异常是解决我们挑战的一个潜在方案。关键问题是:现有的VLMs是否真正理解物理规则并以正确的方式进行推理？专门为视频理解方法设计，我们引入了一个名为物理异常解释(PAEval)的新指标。具体来说，PAEval从三个不同角度评估基于VLMs的算法的异常检测性能:分类、描述和解释。分类指的是传统的异常检测指标，如AUROC等。受 工作的启发，PAEval还引入了两个额外的评估指标:描述和物理解释。描述指的是模型正确描述视频内容的能力，用于评估模型是否具备描述物理现象的能力。解释指的是模型正确解释视频中异常物理原因的能力，用于评估VLM的推理能力。为了为描述和解释指标提供标签，我们手动标注了所有类别中的每种缺陷，并通过ChatGPT进行数据增强，以确保检测的鲁棒性。PAEval指标的整个流程如图6所示。

# 5. Benchmarking Results

# 5. 基准测试结果

# 5.1. Benchmark Methods Selection

# 5.1. 基准方法选择

For the Phys-AD setting, we select popular and reproducible video anomaly detection methods across unsupervised, weakly-supervised, and video-understanding setting. In the unsupervised setting, we focus on reconstruction, prediction, and embedding-based models like MemAE [16] and MNAD [40], which use memory modules to enhance anomaly discrimination. For weakly-supervised anomaly detection, we adopt methods such as VadCLIP [51] and MGFN [10], which use feature magnitudes and vision-language associations. In video-understanding, we evaluate video-language models like Video-ChatGPT [36] and image-language models like CLIP [42] to predict detailed anomaly descriptions and scores through cross-modal embeddings.

对于Phys-AD设置，我们选择了在无监督、弱监督和视频理解设置中流行且可复现的视频异常检测方法。在无监督设置中，我们专注于基于重建、预测和嵌入的模型，如MemAE [16]和MNAD [40]，它们使用记忆模块来增强异常判别能力。对于弱监督异常检测，我们采用了如VadCLIP [51]和MGFN [10]等方法，它们使用特征幅度和视觉语言关联。在视频理解中，我们评估了如Video-ChatGPT [36]等视频语言模型和如CLIP [42]等图像语言模型，通过跨模态嵌入预测详细的异常描述和分数。

Table 5. Physics Anomaly Explanation (PAEval) metric results on our Phys-AD dataset.

表5. 在我们的Phys-AD数据集上的物理异常解释(PAEval)指标结果。

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | Classification (↑) | Description (↑) | Explanation (↑) |
| LAVAD[56] | 0.510 | 0.157 | 0.000 |
| Video-ChatGPT[36] | 0.496 | 0.131 | 0.160 |
| Video-LLaMA[59] | 0.523 | 0.137 | 0.303 |
| Video-LLaVA[28] | 0.463 | 0.219 | 0.282 |

|  |  |  |  |
| --- | --- | --- | --- |
| 方法 | 分类 (↑) | 描述 (↑) | 解释 (↑) |
| LAVAD[56] | 0.510 | 0.157 | 0.000 |
| Video-ChatGPT[36] | 0.496 | 0.131 | 0.160 |
| Video-LLaMA[59] | 0.523 | 0.137 | 0.303 |
| Video-LLaVA[28] | 0.463 | 0.219 | 0.282 |

Code and Experiment details. We provide code and toolkit for our dataset and benchmark. More experiment details are listed in Appendix.

代码和实验细节。我们提供了数据集和基准测试的代码和工具包。更多实验细节列在附录中。

# 5.2. Main Findings

# 5.2. 主要发现

Overall anomaly detection benchmarking results. Table 4 shows that existing video anomaly detection and video-understanding methods achieve limited performance on the Phys-AD dataset, with the highest AUROC only reaching for MNAD.r [40]. This result underscores the heightened complexity of Phys-AD compared to existing datasets, highlighting a domain shift that challenges current industrial anomaly detection algorithms, which are often tuned to visually distinct anomaly patterns in single frames rather than complex temporal or physical cues.

整体异常检测基准测试结果。表4显示，现有的视频异常检测和视频理解方法在Phys-AD数据集上表现有限，MNAD.r [40]的最高AUROC仅达到 。这一结果突显了Phys-AD与现有数据集相比的更高复杂性，强调了领域转移对当前工业异常检测算法的挑战，这些算法通常针对单帧中视觉上明显的异常模式进行调整，而不是复杂的时间或物理线索。

Anomaly detection via unsupervised and weakly supervised methods. Our experiment includes unsupervised methods (e.g., MemAE [16], MPN [34], MNAD [40]) and weakly supervised methods (e.g., S3R [48], MGFN [10], VAD-Clip [51]). Unsupervised methods like MNAD.p [40] performed better on temporal anomalies (81.0% for fan, 68.0% for screw) by leveraging prediction-based approaches, which excel in scenarios requiring temporal understanding. Weakly supervised methods improve baseline scores across complex classes, preventing extreme low scores in challenging categories such as spherical bearing. However, weakly supervised methods show marginally lower performance in simpler anomaly classes, indicating trade-offs introduced by anomaly samples in training.

通过无监督和弱监督方法进行异常检测。我们的实验包括无监督方法(如MemAE [16]、MPN [34]、MNAD [40])和弱监督方法(如S3R [48]、MGFN [10]、VAD-Clip [51])。像MNAD.p [40]这样的无监督方法通过利用基于预测的方法在时间异常上表现更好(风扇为81.0%，螺丝为68.0%)，这些方法在需要时间理解的场景中表现出色。弱监督方法在复杂类别中提高了基线分数，防止了在挑战性类别(如球形轴承)中出现极低分数。然而，弱监督方法在较简单的异常类别中表现略低，表明训练中异常样本引入的权衡。

Unsupervised anomaly detection method performance by category. Among unsupervised methods, MemAE [16] achieves high AUROC on objects with spatial anomalies (e.g., sticky roller, 96.7%) but underperforms in temporal anomaly classes like rolling and spherical bearings, where it achieves only 4.4% and 9.2%, respectively. MNAD [40] improves temporal sensitivity by incorporating temporal prototypes, achieving better scores for bearings but still struggling with complex anomaly types, revealing limitations in purely prototype-based approaches.

按类别划分的无监督异常检测方法性能。在无监督方法中，MemAE [16]在具有空间异常的对象上实现了高AUROC(如粘性滚筒，96.7%)，但在滚动和球形轴承等时间异常类别中表现不佳，分别仅为4.4%和9.2%。MNAD [40]通过引入时间原型提高了时间敏感性，在轴承上获得了更好的分数，但在复杂异常类型上仍然表现不佳，揭示了纯基于原型的方法的局限性。

Anomaly detection via Video-Language Models. VLM-based methods (e.g., Video-ChatGPT [36], Video-LLaMA [59]) struggle, with top-performing Video-LLaMA [59] reaching only average AUROC. Performance is impacted by reliance on pre-trained weights that are not optimized for physics-grounded video content, evident from low AUROC scores in categories with nuanced physical dynamics, such as hinges and screws. Further, PAEval results suggest that these models lack effective reasoning about object physical dynamics and behaviors influenced by physical forces, underscoring a gap between VLM capabilities and the demands of IAD tasks.

通过视频语言模型进行异常检测。基于VLM的方法(如Video-ChatGPT [36]、Video-LLaMA [59])表现不佳，表现最好的Video-LLaMA [59]平均AUROC仅为 。性能受到依赖未针对物理基础视频内容优化的预训练权重的影响，这在具有细微物理动态的类别(如铰链和螺丝)中表现明显。此外，PAEval结果表明，这些模型缺乏对受物理力影响的物体物理动态和行为的有效推理，突显了VLM能力与IAD任务需求之间的差距。

Anomaly explanation via Video-Language Models. Table 5 reports PAEval metric results, with the best-performing VLM achieving only in description and in explanation. These findings emphasize that current VLMs lack the depth in physical reasoning and temporal coherence required for understanding real-world physics-based scenarios, which our Phys-AD dataset demands.

通过视频语言模型进行异常解释。表5报告了PAEval指标结果，表现最好的VLM在描述和解释中分别仅达到 和 。这些发现强调，当前的VLM缺乏理解现实世界物理场景所需的物理推理和时间一致性深度，而我们的Phys-AD数据集要求这些能力。

Summary. Overall, the results reveal Phys-AD’s unique challenge in requiring both high spatial detail and temporal comprehension, areas where existing methods and models underperform. This analysis points to the need for future research in models that integrate temporal reasoning with physics-based anomaly detection.

总结。总体而言，结果揭示了Phys-AD在需要高空间细节和时间理解方面的独特挑战，现有方法和模型在这些方面表现不佳。这一分析指出，未来需要在将时间推理与基于物理的异常检测相结合的模型中进行研究。

# 6. Conclusion

# 6. 结论

In this paper, we introduce the first industrial anomaly detection task focusing on real-world scenarios where physical understanding and reasoning are essential for anomaly detection. We present the Physics Anomaly Detection (Phys-AD) dataset, a large-scale, physics-grounded video dataset with over 6400 videos across 22 categories and 49 object types interacting with robotic systems, capturing 47 anomaly types that necessitate visual and physical understanding. We assess Phys-AD, highlighting the lack of baseline approaches for high-level reasoning in anomaly detection. Additionally, we propose the Physics Anomaly Explanation (PAE-val) metric to evaluate visual language models (VLMs) on physics-based reasoning. Experiments show that current VLMs fall short of human-level understanding in physics-based anomaly scenarios. This work marks a milestone for industrial anomaly detection, promoting physics-grounded reasoning in complex industrial settings.

在本文中，我们介绍了第一个专注于现实世界场景的工业异常检测任务，其中物理理解和推理对异常检测至关重要。我们提出了物理异常检测(Phys-AD)数据集，这是一个大规模的、基于物理的视频数据集，包含6400多个视频，涵盖22个类别和49种与机器人系统交互的对象类型，捕捉了47种需要视觉和物理理解的异常类型。我们评估了Phys-AD，强调了在异常检测中缺乏高级推理的基线方法。此外，我们提出了物理异常解释(PAE-val)指标，以评估视觉语言模型(VLM)在基于物理的推理上的表现。实验表明，当前的VLM在基于物理的异常场景中未能达到人类理解水平。这项工作标志着工业异常检测的一个里程碑，促进了复杂工业环境中基于物理的推理。

Limitation and future work. Although our Phys-AD dataset provides a large variety of objects with diverse physical properties and various types of interaction methods, we plan to add even more diverse interaction methods and objects in the future to better meet the demands of complex real-world industrial scenarios. Due to the significant differences between our Phys-AD dataset and current industrial anomaly detection and video anomaly detection datasets, most existing anomaly detection algorithms cannot be directly applied to our dataset. In the future, we will test more algorithms on our Phys-AD dataset and provide experimental results across more settings like zero-shot, few-shot, semi-supervised settings, etc.

局限性和未来工作。尽管我们的Phys-AD数据集提供了具有多种物理属性和各种交互方法的大量对象，我们计划在未来添加更多样化的交互方法和对象，以更好地满足复杂现实世界工业场景的需求。由于我们的Phys-AD数据集与当前工业异常检测和视频异常检测数据集之间存在显著差异，大多数现有的异常检测算法无法直接应用于我们的数据集。未来，我们将在Phys-AD数据集上测试更多算法，并提供更多设置(如零样本、少样本、半监督设置等)的实验结果。

References

参考文献

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In

斯坦尼斯瓦夫·安托尔、艾什瓦里亚·阿格拉瓦尔、贾森·卢、玛格丽特·米切尔、德鲁夫·巴特拉、C·劳伦斯·齐特尼克和德维·帕里克。VQA:视觉问答。在 中

[2] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. arXiv preprint arXiv:1909.12000, 2019. 3

法比安·巴拉德尔、娜塔莉亚·内维罗娃、朱利安·米勒、格雷格·莫里和克里斯蒂安·沃尔夫。Cophy:物理动态的反事实学习。arXiv预印本arXiv:1909.12000，2019年。3

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9584-9592, 2019. 3

保罗·伯格曼、迈克尔·福瑟、大卫·萨特勒格和卡斯滕·斯特格。Mvtec AD — 一个用于无监督异常检测的全面真实世界数据集。在2019年IEEE/CVF计算机视觉与模式识别会议(CVPR)上，第9584-9592页，2019年。3

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In CVPR, 2019. 2

保罗·伯格曼、迈克尔·福瑟、大卫·萨特勒格和卡斯滕·斯特格。MVTec AD - 一个用于无监督异常检测的全面真实世界数据集。在CVPR，2019年。2

[5] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sat-tlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. International Journal of Computer Vision, 130 (4):947-969, 2022. 3

保罗·伯格曼、基利安·巴茨纳、迈克尔·福瑟、大卫·萨特勒格和卡斯滕·斯特格。超越凹痕和划痕:无监督异常检测和定位中的逻辑约束。国际计算机视觉杂志，130(4):947-969，2022年。3

[6] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec -ad dataset for unsupervised anomaly detection and localization. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS - Science and Technology Publications, 2022. 3

保罗·伯格曼、辛金、大卫·萨特勒格和卡斯滕·斯特格。用于无监督 异常检测和定位的mvtec -ad数据集。在第17届国际计算机视觉、成像和计算机图形理论与应用联合会议论文集上。SCITEPRESS - 科学与技术出版物，2022年。3

[7] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3d-AD dataset for unsupervised 3d anomaly detection and localization. In VISAPP, 2022. 2

保罗·伯格曼、辛金、大卫·萨特勒格和卡斯滕·斯特格。用于无监督3D异常检测和定位的MVTec 3d-AD数据集。在VISAPP，2022年。2

[8] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In Proceedings of the Asian Conference on Computer Vision, pages 3586-3602, 2022. 3

卢卡·邦菲廖利、马尔科·托斯基、达维德·西尔维斯特里、尼古拉·菲奥拉约和达尼埃莱·德·格雷戈里奥。用于无监督多模态异常检测和定位的eyecandies数据集。在亚洲计算机视觉会议论文集上，第3586-3602页，2022年。3

[9] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. arXiv preprint arXiv:2401.16402, 2024. 2

曹云康、徐晓浩、张江宁、程宇琪、黄晓楠、庞冠松和沈伟明。视觉异常检测综述:挑战、方法与前景。arXiv预印本arXiv:2401.16402，2024年。2

[10] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In Proceedings of the AAAI conference on artificial intelligence, pages 387-395, 2023. 7, 8

陈英贤、刘正哲、张宝恒、霍伟顿、齐晓娟和吴奕聪。Mgfn:用于弱监督视频异常检测的幅度对比瞥视与聚焦网络。在AAAI人工智能会议论文集上，第387-395页，2023年。7, 8

[11] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18793- 18803, 2024. 7

杜航、张思成、谢斌珠、南国顺、张佳阳、徐俊瑞、刘航宇、冷思聪、刘江明、范赫赫等。揭示什么、为什么和如何:视频异常因果理解的全面基准。在IEEE/CVF计算机视觉与模式识别会议论文集上，第18793-18803页，2024年。7

[12] Jiafei Duan, Samson Yu, and Cheston Tan. Space: A simulator for physical interactions and causal learning in environments. In , pages

段佳飞、余山姆森和陈志森。Space:一个用于 环境中物理交互和因果学习的模拟器。在 中，第 页

[13] Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. Pip: Physical interaction prediction via mental

段佳飞、余山姆森、苏贾尼亚·波里亚、温碧涵和陈志森。Pip:通过心理模拟与跨度选择的物理交互预测

simulation with span selection. In , pages 405-421.

在第 页，第405-421页。

Springer, 2022. 3

Springer，2022年。3

[14] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14009-14018, 2021. 3

冯家昌、洪发廷和郑伟石。Mist:用于视频异常检测的多实例自训练框架。在IEEE/CVF计算机视觉与模式识别会议论文集上，第14009-14018页，2021年。3

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15180-15190, 2023. 7

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, 和 Ishan Misra。Imagebind:一个嵌入空间绑定所有。在IEEE/CVF计算机视觉与模式识别会议论文集，第15180-15190页，2023年。7

[16] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1705-1714, 2019. 3, 7, 8

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, 和 Anton van den Hengel。通过记忆正常来检测异常:用于无监督异常检测的记忆增强深度自编码器。在IEEE/CVF国际计算机视觉会议论文集，第1705-1714页，2019年。3, 7, 8

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Ba-tra, and Devi Parikh. Making the in vqa matter: Elevating the role of image understanding in visual question answering. In

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Ba-tra, 和 Devi Parikh。让VQA中的 重要:提升图像理解在视觉问答中的作用。在

[18] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, 和 Larry S Davis。学习视频序列中的时间规律性。在

[19] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In

Drew A. Hudson 和 Christopher D. Manning。Gqa:一个用于现实世界视觉推理和组合问答的新数据集。在

[20] Steeven Janny, Fabien Baradel, Natalia Neverova, Madiha Nadri, Greg Mori, and Christian Wolf. Filtered-cophy: Unsupervised learning of counterfactual physics in pixel space. In ICLR, 2022. 3

Steeven Janny, Fabien Baradel, Natalia Neverova, Madiha Nadri, Greg Mori, 和 Christian Wolf。Filtered-cophy:在像素空间中无监督学习反事实物理。在ICLR，2022年。3

[21] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pages 66-71, 2021. 2, 3

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, 和 Milos Skotak。基于深度学习的金属零件缺陷检测:在复杂条件下评估当前方法。在2021年第13届国际超现代电信与控制系统及研讨会(ICUMT)，第66-71页，2021年。2, 3

[22] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In CVPR, pages 19056-19065, 2022. 3

Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, 和 Anima Anandkumar。Bongard-hoi:用于人-物交互的少样本视觉推理基准。在CVPR，第19056-19065页，2022年。3

[23] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, 和 Ross Girshick。Clevr:一个用于组合语言和基本视觉推理的诊断数据集。在

[24] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In 2023 IEEE International Conference on Image Processing (ICIP), pages 3230- 3234. IEEE, 2023. 3

Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, 和 Ngan Le。Clip-tsa:用于弱监督视频异常检测的Clip辅助时间自注意力。在2023年IEEE国际图像处理会议(ICIP)，第3230-3234页。IEEE，2023年。3

[25] Federico Landi, Cees GM Snoek, and Rita Cucchiara. Anomaly locality in video surveillance. arXiv preprint arXiv:1901.10364, 2019. 3

Federico Landi, Cees GM Snoek, 和 Rita Cucchiara。视频监控中的异常局部性。arXiv预印本arXiv:1901.10364，2019年。3

[26] Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, Donald Geman, et al. Comparing machines and humans on a visual categorization test. 2011. 3

Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, Donald Geman, 等。在视觉分类测试中比较机器和人类。2011年。3

[27] Wenqiao Li, Xiaohao Xu, Yao Gu, Bozhong Zheng, Shenghua Gao, and Yingna Wu. Towards scalable 3d anomaly detection and localization: A benchmark via 3d anomaly synthesis and a self-supervised learning network. arXiv preprint, 2023. 2, 3

Wenqiao Li, Xiaohao Xu, Yao Gu, Bozhong Zheng, Shenghua Gao, 和 Yingna Wu。迈向可扩展的3D异常检测与定位:通过3D异常合成和自监督学习网络的基准。arXiv预印本，2023年。2, 3

[28] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122,2023.7,8

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, 和 Li Yuan。Video-llava:通过投影前的对齐学习统一的视觉表示。arXiv预印本arXiv:2311.10122，2023年。7,8

[29] Jiaqi Liu, Guoyang Xie, Rui Chen, et al. Real3d-ad: A dataset of point cloud anomaly detection. NeurIPS, 2023. 2

Jiaqi Liu, Guoyang Xie, Rui Chen, 等。Real3d-ad:一个点云异常检测的数据集。NeurIPS，2023年。2

[30] Jiaqi Liu, Guoyang Xie, Ruitao Chen, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, and Feng Zheng. Real3dad: A dataset of point cloud anomaly detection. Advances in Neural Information Processing Systems, 36, 2024. 3

刘佳琪、谢国阳、陈瑞涛、李新鹏、王金宝、刘勇、王成杰和郑峰。Real3dad:点云异常检测数据集。神经信息处理系统进展，36，2024年。3

[31] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In Proceedings of the 27th ACM International Conference on Multimedia, pages 1490-1499, 2019. 3

刘坤和马华东。探索背景偏差用于监控视频中的异常检测。第27届ACM国际多媒体会议论文集，1490-1499页，2019年。3

[32] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection-a new baseline. In

刘文、罗伟新、连东泽和高升华。未来帧预测用于异常检测——新基线。

[33] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In Proceedings of the IEEE/CVF international conference on computer vision, pages 13588-13597, 2021. 3

刘志安、聂永伟、龙成江、张庆和李桂清。通过记忆增强流重建和流引导帧预测的混合视频异常检测框架。IEEE/CVF国际计算机视觉会议论文集，13588-13597页，2021年。3

[34] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15425-15434, 2021. 7, 8

吕辉、陈晨、崔振、徐春燕、李勇和杨健。通过元原型网络学习视频中的正常动态。IEEE/CVF计算机视觉与模式识别会议论文集，15425-15434页，2021年。7, 8

[35] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8022-8031, 2023. 3

吕辉、岳中奇、孙倩茹、罗斌、崔振和张汉旺。无偏多实例学习用于弱监督视频异常检测。IEEE/CVF计算机视觉与模式识别会议论文集，8022-8031页，2023年。3

[36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023. 7, 8

穆罕默德·马兹、哈努娜·拉希德、萨尔曼·汗和法哈德·沙赫巴兹·汗。Video-chatgpt:通过大规模视觉和语言模型实现详细视频理解。arXiv预印本arXiv:2306.05424，2023年。7, 8

[37] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of x-ray images for nondestructive testing. Journal of Nondestructive Evaluation, 34 (4):42, 2015. 3

多明戈·梅里、弗拉基米尔·里福、乌韦·兹谢佩尔、赫尔曼·蒙德拉贡、伊万·利略、伊雷内·祖卡尔、汉斯·洛贝尔和米格尔·卡拉斯科。Gdxray:无损检测X射线图像数据库。无损评估杂志，34(4):42，2015年。3

[38] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE). IEEE, 2021. 3

潘卡吉·米什拉、里卡多·维尔克、达尼埃莱·福纳西尔、克劳迪奥·皮恰雷利和吉安·卢卡·福雷斯蒂。Vt-adl:用于图像异常检测和定位的视觉Transformer网络。2021年IEEE第30届国际工业电子研讨会(ISIE)。IEEE，2021年。3

[39] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Animashree Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. In NeurIPS, 2020. 3

魏立·聂、于志定、毛磊、安基特·B·帕特尔、朱宇克和阿尼玛什里·阿南德库马尔。Bongard-logo:人类水平概念学习和推理的新基准。NeurIPS，2020年。3

[40] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In Proceed-

朴贤钟、卢钟允和咸范燮。学习记忆引导的正常性用于异常检测。IEEE/CVF计算机视觉与模式识别会议论文集，14372-14381页，2020年。7, 8

ings of the IEEE/CVF conference on computer vision and

IEEE/CVF计算机视觉与模式识别会议论文集，14372-14381页，2020年。7, 8

pattern recognition, pages 14372-14381, 2020. 7, 8

IEEE/CVF计算机视觉与模式识别会议论文集，14372-14381页，2020年。7, 8

[41] Yujiang Pu, Xiaoyu Wu, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. arXiv preprint arXiv:2306.14451, 2023. 3

蒲玉江、吴晓宇和王升金。学习提示增强的上下文特征用于弱监督视频异常检测。arXiv预印本arXiv:2306.14451，2023年。3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748-8763. PMLR, 2021. 3, 7

亚历克·拉德福德、金钟旭、克里斯·哈拉西、阿迪亚·拉梅什、高加布里埃尔、桑迪尼·阿加瓦尔、吉里什·萨斯特里、阿曼达·阿斯克尔、帕梅拉·米什金、杰克·克拉克等。从自然语言监督中学习可迁移的视觉模型。国际机器学习会议，8748-8763页。PMLR，2021年。3, 7

[43] Binyi Su, Zhong Zhou, and Haiyong Chen. Pvel-ad: A large-scale open-world dataset for photovoltaic cell anomaly detection. IEEE Transactions on Industrial Informatics, 19(1): 404-413, 2023. 3

苏斌毅、周忠和陈海勇。Pvel-ad:用于光伏电池异常检测的大规模开放世界数据集。IEEE工业信息学汇刊，19(1):404-413，2023年。3

[44] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In CVPR, 2018. 7

Waqas Sultani, Chen Chen, 和 Mubarak Shah。监控视频中的真实世界异常检测。在 CVPR 中，2018. 7

[45] Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In Proceedings of the IEEE/CVF Winter conference on applications of computer vision, pages 5541-5550, 2023. 3

Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, 和 Ig-Jae Kim。Dyannet:一种场景动态性引导的自训练视频异常检测网络。在 IEEE/CVF 冬季计算机视觉应用会议论文集，第 5541-5550 页，2023. 3

[46] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jianning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection, 2024. 3

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jianning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, 和 Lizhuang Ma。Real-iad:一个用于基准测试多功能工业异常检测的真实世界多视图数据集，2024. 3

[47] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8032-8041, 2023. 7

Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, 和 Chengjie Wang。通过混合融合进行多模态工业异常检测。在 IEEE/CVF 计算机视觉与模式识别会议论文集，第 8032-8041 页，2023. 7

[48] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In European Conference on Computer Vision, pages 729-745. Springer, 2022. 7, 8

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, 和 Tyng-Luh Liu。自监督稀疏表示用于视频异常检测。在欧洲计算机视觉会议，第 729-745 页。Springer，2022. 7, 8

[49] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX 16, pages 322-339. Springer, 2020. 3, 7

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, 和 Zhiwei Yang。不仅看，还要听:在弱监督下学习多模态暴力检测。在计算机视觉-ECCV 2020:第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23-28 日，论文集，第 XXX 16 部分，第 322-339 页。Springer，2020. 3, 7

[50] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. arXiv preprint arXiv:2311.07042, 2023. 3

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, 和 Yanning Zhang。开放词汇视频异常检测。arXiv 预印本 arXiv:2311.07042，2023. 3

[51] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. arXiv preprint arXiv:2308.11681, 2023.3,7,8

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, 和 Yanning Zhang。Vadclip:适应视觉语言模型用于弱监督视频异常检测。arXiv 预印本 arXiv:2308.11681，2023.3,7,8

[52] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 156:117-127, 2017. 3

Dan Xu, Yan Yan, Elisa Ricci, 和 Nicu Sebe。通过学习外观和运动的深度表示来检测视频中的异常事件。计算机视觉与图像理解，156:117-127，2017. 3

[53] Xiaohao Xu, Yunkang Cao, Yongqi Chen, Weiming Shen, and Xiaonan Huang. Customizing visual-language foundation

Xiaohao Xu, Yunkang Cao, Yongqi Chen, Weiming Shen, 和 Xiaonan Huang。定制视觉语言基础

models for multi-modal anomaly detection and reasoning. arXiv preprint arXiv:2403.11083, 2024. 3

模型用于多模态异常检测和推理。arXiv 预印本 arXiv:2403.11083，2024. 3

[54] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, 和 Xiaotao Liu。基于关键帧的视频事件恢复用于视频异常检测。在

[55] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14744-14754, 2022. 3

M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, 和 Seung-Ik Lee。生成式协作学习用于无监督视频异常检测。在 IEEE/CVF 计算机视觉与模式识别会议论文集，第 14744-14754 页，2022. 3

[56] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yim-ing Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. arXiv preprint arXiv:2404.01014, 2024. 3, 7, 8

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yim-ing Wang, 和 Elisa Ricci。利用大语言模型进行无训练视频异常检测。arXiv 预印本 arXiv:2404.01014，2024. 3, 7, 8

[57] Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Mus-slick, and Thomas Serre. A benchmark for compositional visual reasoning. arXiv preprint arXiv:2206.05379, 2022. 3

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Mus-slick, 和 Thomas Serre。组合视觉推理的基准。arXiv 预印本 arXiv:2206.05379，2022. 3

[58] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, 和 Song-Chun Zhu。Raven:一个用于关系和类比视觉推理的数据集。在

[59] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023. 7, 8

张航, 李鑫, 和宾立东. Video-llama: 一种用于视频理解的指令调优视听语言模型. arXiv预印本 arXiv:2306.02858, 2023. 7, 8

[60] Huaxin Zhang, Xiang Wang, Xiaohao Xu, Xiaonan Huang, Chuchu Han, Yuehuan Wang, Changxin Gao, Shanjun Zhang, and Nong Sang. Glancevad: Exploring glance supervision for label-efficient video anomaly detection. arXiv preprint arXiv:2403.06154, 2024. 3

张华新, 王翔, 徐晓浩, 黄晓楠, 韩楚楚, 王月环, 高长新, 张善军, 和桑农. Glancevad: 探索用于标签高效视频异常检测的瞥见监督. arXiv预印本 arXiv:2403.06154, 2024. 3

[61] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. arXiv preprint arXiv:2406.12235, 2024. 3, 7

张华新, 徐晓浩, 王翔, 左家龙, 韩楚楚, 黄晓楠, 高长新, 王月环, 和桑农. Holmes-vad: 通过多模态LLM实现无偏见和可解释的视频异常检测. arXiv预印本 arXiv:2406.12235, 2024. 3, 7

[62] Yiheng Zhang, Yunkang Cao, Xiaohao Xu, and Weiming Shen. Logicode: An llm-driven framework for logical anomaly detection. IEEE Transactions on Automation Science and Engineering, pages 1-0, 2024. 3

张一恒, 曹云康, 徐晓浩, 和沈伟明. Logicode: 一种LLM驱动的逻辑异常检测框架. IEEE自动化科学与工程学报, 页码 1-0, 2024. 3

[63] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In

钟家兴, 李楠楠, 孔伟杰, 刘山, 李托马斯, 和李戈. 图卷积标签噪声清理器: 训练一个即插即用的动作分类器用于异常检测. 在

[64] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad: A dataset and benchmark for pose-agnostic anomaly detection, 2023. 3

周强, 李伟泽, 蒋立涵, 王国良, 周古月, 张尚航, 和赵浩. Pad: 一个用于姿态无关异常检测的数据集和基准, 2023. 3

[65] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In , 2022. 2

邹阳, 郑钟赫, 佩穆拉·拉塔, 张东青, 和达比尔·昂卡尔. 用于异常检测和分割的找不同自监督预训练. 在 , 2022. 2

[66] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation, 2022. 3

邹阳, 郑钟赫, 佩穆拉·拉塔, 张东青, 和达比尔·昂卡尔. 用于异常检测和分割的找不同自监督预训练, 2022. 3

1. \*Equally contribute to this work.

   \*对本工作有同等贡献。

   Corresponding authors

   通讯作者 [↑](#footnote-ref-27)