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

Wenqiao Li[1*] Yao Gu[1*] Xintao Chen[1*] Xiaohao Xu[2†] Ming Hu[3] Xiaonan Huang[2] Yingna Wu[1†]
[1]ShanghaiTech University [2]University of Michigan, Ann Arbor [3]Monash University

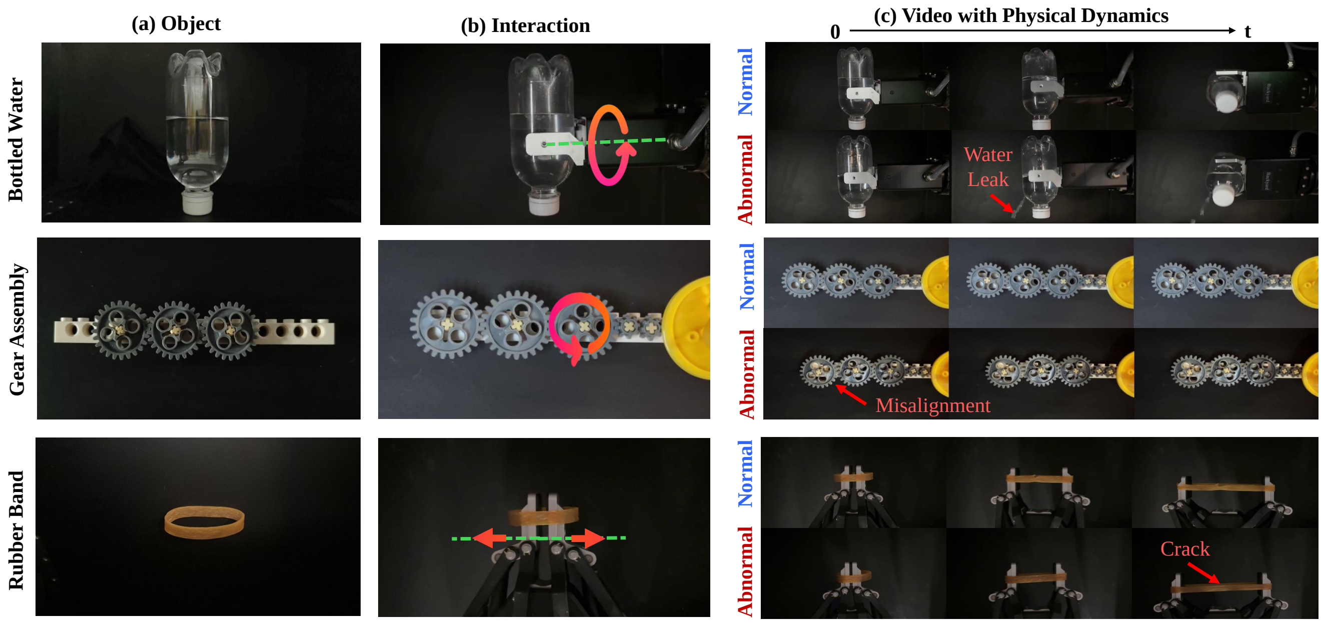

Figure 1. **Towards visual discrimination of physical dynamics in real-world industrial object anomaly detection.** We illustrate objects, interactions, and time-sequenced videos from the ***Physics-Grounded Anomaly Detection*** dataset: **(a) Object**; **(b) Interaction:** Applied actions shown with 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.

Abstract

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 AD, 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.

*Equally contribute to this work
†Corresponding authors

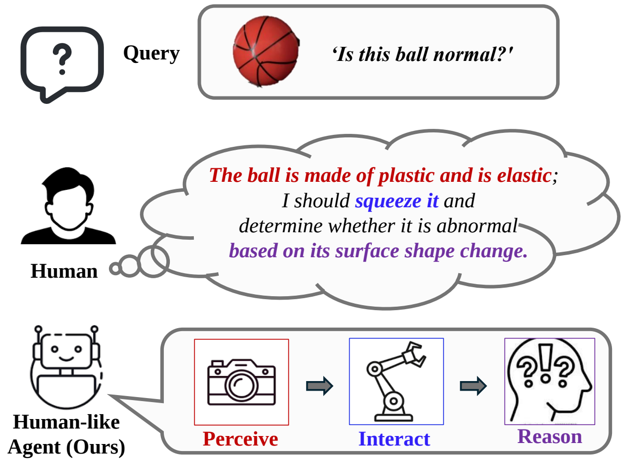

Figure 2. **Human-like decision-making process for physics-grounded object anomaly detection.** We illustrate the sequential approach of a human-like agent for evaluating an object's normality. First, the agent perceives relevant physical attributes (e.g., plastic and elastic), then interacts by performing 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.

1. Introduction

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.

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 [66] 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.

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.

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.

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.

Our contributions can be summarized as followings:

- We introduce a novel task of detecting physical-based industrial anomalies in real-world that involves perception, physical and visual reasoning.
- 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.
- We benchmarking the anomaly detection and reasoning

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.

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 [67]	2021	Real	RGB	12	-	✗
MVTec LOCO-AD [5]	2022	Real	RGB	5	-	✗
MAD [65]	2023	Syn+Real	RGB	20	3	✗
LOCO-Annotations [63]	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**	✓

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.

2. Related work

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 [67] 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] proposed in 2022, 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 [65] 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.

Video anomaly detection. Deep learning methods [18, 35, 55, 62] now dominate Video Anomaly Detection (VAD), categorized into unsupervised, weakly-supervised, and fully-supervised approaches. Unsupervised methods learn normal patterns via reconstruction [16, 18, 53], prediction [32], or hybrids [33], while some methods [45, 56] train with both unlabeled normal and abnormal data. Weakly-supervised methods[14, 50, 61, 64] 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 [24, 41, 52], focusing on semantic anomalies. Open-vocabulary VAD [51] and prompt-based anomaly scoring [57] leverage LLMs [54, 62], 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.

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 [1, 17, 19, 23]. For 2D puzzles, tasks involve discovering relationships among visual elements and making inferences [22, 26, 39, 58, 59]. 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***.

3. Dataset: Phys-AD

3.1. Object Preparation and Interaction Selection

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

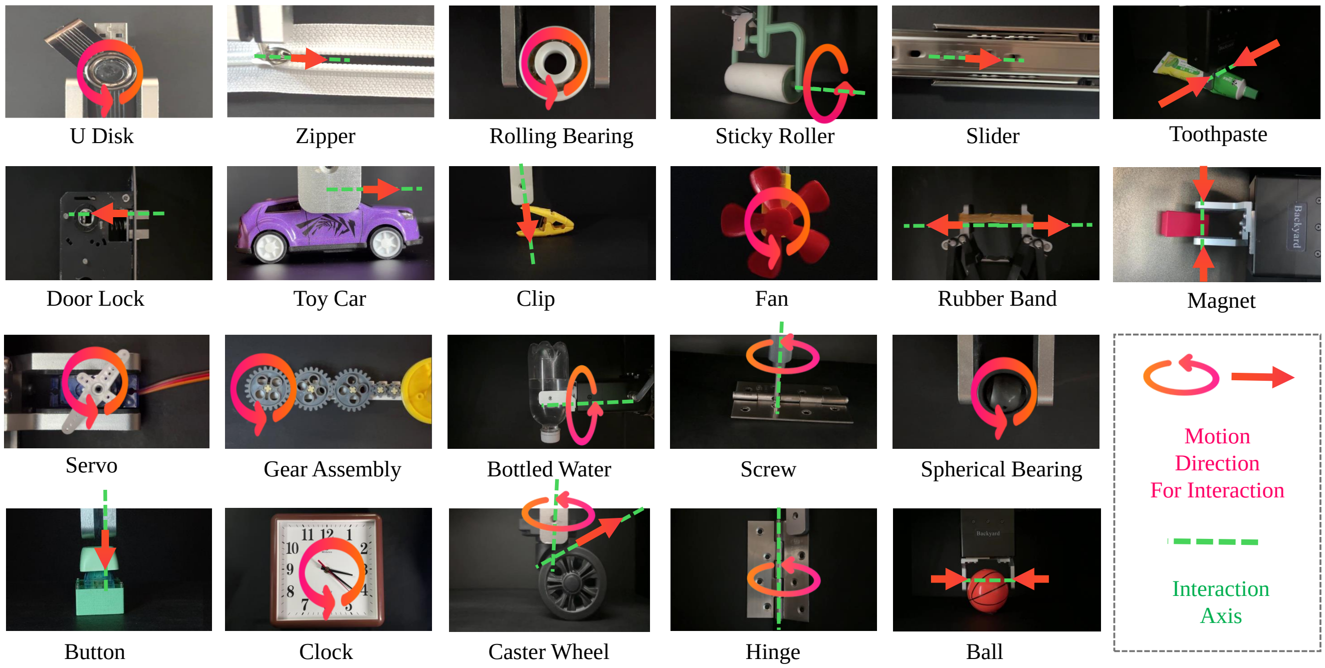

Figure 3. **Interactions for understanding implicit physical laws in the Phys-AD dataset.** We showcase various object interactions from the Phys-AD dataset, where different actions (indicated by motion directions) are used to explore and reason about the underlying physical properties and behaviors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

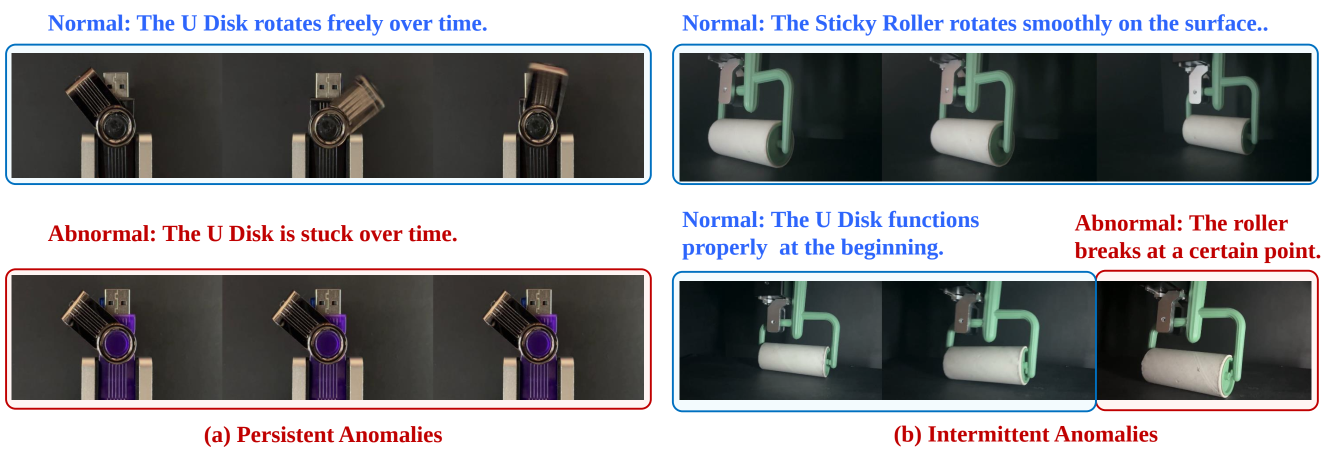

Figure 4. **Categorization of anomalies based on persistence in the Phys-AD dataset.** We show examples of normal and abnormal functioning in common objects, divided into two anomaly types: persistent and intermittent. (**a**) Persistent anomalies, such as continuous obstruction in the U Disk or permanent 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.

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 USB drive 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.

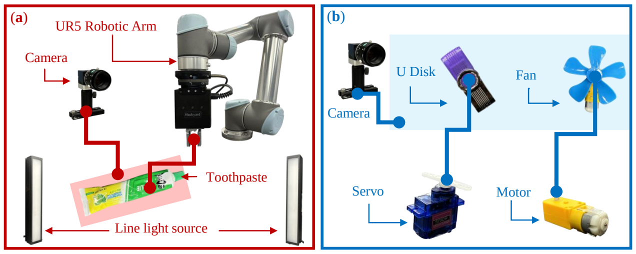

Figure 5. **Data collection pipeline for the Phys-AD dataset.** (**a**) Manipulation of a toothpaste tube using a UR5 robotic arm. (**b**) Manipulation of a U Disk and fan via servo and motor.

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.

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
USB	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

3.2. Data Collection and Processing

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 1080p resolution and a frame rate of 60 FPS, providing high-quality video sequences. Some kinds of the objects like U Disk that are not suitable for robotics arms' manipulation, 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.

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.

Category	Train #Frames	Train #Videos	Test #Frames	Test #Videos	Test #Videos Nor.	Test #Videos Ab.	Obj. Types	Def. 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
USB	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

3.3. Data Statistics

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.

Interaction and defect types across categories. Table 2 outlines the specific defect types and corresponding interaction methods required for each category in the Phys-AD dataset. The dataset covers a wide range of physical objects, each with distinct interaction methods and anomaly types. To be more specific, Phys-AD dataset contains various interaction methods like Rotate or Grab and 47 defect types across 22 categories. It is worth noting that the interaction methods or defect types with the same name vary between different categories. For example, the rotation of the fan refers to the rotation of the blades driven by the motor, while the rotation of the hinge represents the rotation of the hinge's pages in the direction of the shaft under the manipulation of the robotic arm. In addition to this, in order to detect anomalies with complex physical properties in a human-like manners, we chose to combine multiple interactions together. For ex-

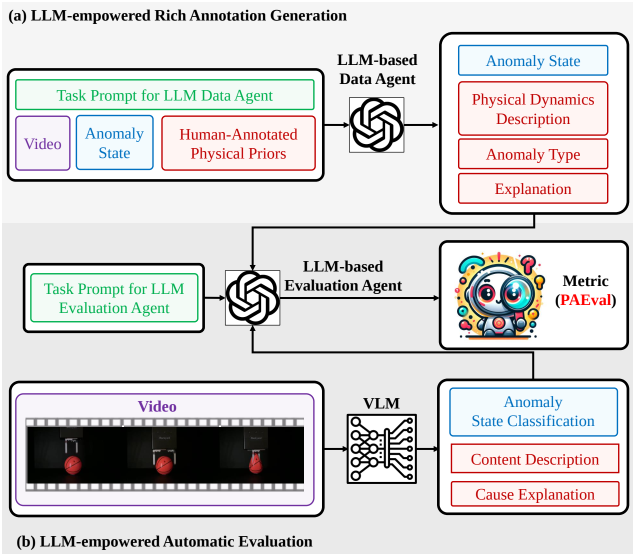

Figure 6. **PhysAD-Agent: A Large Language Model (LLM)-powered system for physics anomaly detection label augmentation and automatic evaluation.** This agent framework consists of two main components: (**a**) ***Rich Annotation Generation***, where an LLM-based data agent generates detailed anomaly annotations 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.

ample, when inspecting the screw, we first press the screw into the hole, and then use the robotic arm to hold the motor to rotate the screw into the deep of the hole. To sum up, Phys-AD dataset provides a diverse and challenging physics-grounded industrial anomaly detection setting, which drive the development of new anomaly detection methods to deal with high-level problem in the real-world.

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.

4. Phys-AD Benchmark

4.1. Problem Definition and Challenges

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

We formulate our Phys-AD setting as two steps.

Step 1: Rules deduction Given a set of training objects $\mathcal{T} = \{t_i\}_{i=1}^{N}$ from category c_i, we first use robotics arms and motors to interact with $\mathcal{T}$ in corresponding methods I_i, and we get the interaction process in video format as V_i. Then, we feed the video sequence V_i and the corresponding category's physical prior P_i together to the deduction function f to obtain the normal rules r_i. After rules deduction for all the categories, we get the rules bank $\mathcal{M} = \{r_i\}_{i=1}^{N}$. Step 1 can be formulated as below:

$$\mathcal{T}\Theta I_i = V_i \tag{1}$$

$$f(V_i|P_i) = r_i \tag{2}$$

$$\mathcal{M} = \{r_i\}_{i=1}^{N} \tag{3}$$

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

Step 2: Anomaly reasoning During test time, we first use robotics arms and motors to interact with $\mathcal{T}$ from the test set in corresponding methods I_i, and we get the interaction process in video format as V_i. Then, we feed the video sequence V_i and rules r_i from step 1 to the reasoner R to predict the object's anomaly score S. Step 2 can be formulated as:

$$\mathcal{T}\Theta I_i = V_i \tag{4}$$

$$R(V_i|r_i) = S, r_i \in \mathcal{M} \tag{5}$$

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.

4.2. Evaluation Settings

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.

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, 2~4 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

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**.

Category.	**Unsupervised**					**Weakly-supervised**			**Video-understanding**					
	MPN [34]	MemAE [16]	MNAD.p [40]	MNAD.r [40]	SVM [44]	VADClip [52]	S3R [48]	MGFN [10]	LAVAD [57]	ZS Clip [42]	ZS ImageBind [15]	Video-ChatGPT [36]	Video-LLaMA [60]	Video-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
USB	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

most anomalies in our test set require reasoning based on the entire video content and physical knowledge of the objects.

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.

4.3. Evaluation Metrics

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].

Physics Anomaly Explanation (PAEval) metric. Directly utilizing existing VLMs [60][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 [11, 62], 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.

5. Benchmarking Results

5.1. Benchmark Methods Selection

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 [52] 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

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

Methods	Classification (↑)	Description (↑)	Explanation (↑)
LAVAD[57]	0.510	0.157	0.000
Video-ChatGPT[36]	0.496	0.131	0.160
Video-LLaMA[60]	**0.523**	0.137	**0.303**
Video-LLaVA[28]	0.463	**0.219**	0.282

descriptions and scores through cross-modal embeddings.

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

5.2. Main Findings

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 66.9% 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.

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 [52]). 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.

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.

Anomaly detection via Video-Language Models. VLM-based methods (e.g., Video-ChatGPT [36], Video-LLaMA [60]) struggle, with top-performing Video-LLaMA [60] reaching only 52.3% 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.

Anomaly explanation via Video-Language Models. Table 5 reports PAEval metric results, with the best-performing VLM achieving only 21.9% in description and 30.3% 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.

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.

6. Conclusion

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 (PAEval) 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.

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*.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 3

[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

[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

[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

[5] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, 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

[6] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d 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

[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

[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

[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

[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

[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

[12] Jiafei Duan, Samson Yu, and Cheston Tan. Space: A simulator for physical interactions and causal learning in 3d environments. In *ICCV*, pages 2058–2063, 2021. 3

[13] Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. Pip: Physical interaction prediction via mental simulation with span selection. In *ECCV*, pages 405–421. Springer, 2022. 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

[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

[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

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 3

[18] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *CVPR*, 2016. 3

[19] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 3

[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

[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

[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

[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 *CVPR*, 2017. 3

[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

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

[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

[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

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

[30] Jiaqi Liu, Guoyang Xie, Ruitao Chen, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, and Feng Zheng. Real3d-ad: A dataset of point cloud anomaly detection. *Advances in Neural Information Processing Systems*, 36, 2024. 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

[32] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, 2018. 3

[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

[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

[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

[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

[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

[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

[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

[40] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 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

[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

[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

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

[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

[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

[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

[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. 7

[50] 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

[51] 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

[52] 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

[53] 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

[54] Xiaohao Xu, Yunkang Cao, Yongqi Chen, Weiming Shen, and Xiaonan Huang. Customizing visual-language foundation models for multi-modal anomaly detection and reasoning. *arXiv preprint arXiv:2403.11083*, 2024. 3

[55] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, 2023. 3

[56] 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

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

[58] Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *arXiv preprint arXiv:2206.05379*, 2022. 3

[59] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, 2019. 3

[60] 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

[61] 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

[62] 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

[63] 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

[64] 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 *CVPR*, 2019. 3

[65] 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

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

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