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

Supplementary Material

In this supplementary material, we provide additional details on our benchmark methods, an in-depth benchmarking analysis, and further illustrations of the Phys-AD dataset with accompanying figures.

A. More Benchmark Methods Details

We provide comprehensive details on the benchmark methods used in our experiments. The methods are grouped into three categories: unsupervised anomaly detection, weakly supervised anomaly detection, and video-understanding-based methods.

A.1. Unsupervised Anomaly Detection Methods

We selected popular and reproducible video anomaly detection algorithms for the unsupervised setting, including reconstruction-based, prediction-based, and embedding-based methods.

A.1.1. MemAE [3]

MemAE is a pioneering work that introduces a memory module to an autoencoder for video frame reconstruction-based anomaly detection methods. It addresses the problem where autoencoders can sometimes reconstruct anomalous parts of the input.

A.1.2. MPN [5]

Based on MemAE, MPN proposes a Dynamic Prototype Unit (DPU) to encode normal dynamics as prototypes in real-time, eliminating extra memory costs.

A.1.3. MNAD [7]

MNAD uses a memory module with a novel update scheme where items in the memory record prototypical patterns of normal data. It presents feature compactness and separateness losses to train the memory, enhancing the discriminative power of both memory items and learned features from normal data. We designed two experimental versions:

- **MNAD.r**: Aimed at current frame reconstruction.
- **MNAD.p**: Aimed at future frame prediction.

A.1.4. SVM [9]

SVM extracts normal temporal features using a pre-trained I3D feature extractor and trains a Support Vector Machine to classify normal and abnormal features.

A.2. Weakly Supervised AD Methods

For the weakly supervised setting, we selected popular methods based on self-supervised learning, feature magnitudes, and clip-based approaches.

A.2.1. S3R [10]

S3R models the feature distribution of normal and abnormal data by combining self-supervised learning with dictionary learning. The training features are used to train a binary classifier for anomaly detection.

A.2.2. MGFN [1]

MGFN proposes the Feature Amplification Mechanism to enhance the discriminativeness of feature magnitudes for anomaly detection.

A.2.3. VAD-CLIP [11]

VAD-CLIP leverages detailed associations between vision and language powered by CLIP and incorporates a dual-branch classifier for anomaly detection.

A.3. Video Understanding Methods

We also evaluated video-language models and image-language models for video understanding and anomaly detection.

A.3.1. Video-LLaMA [13], Video-ChatGPT [6], Video-LLaVA [4]

For these models, we directly fed designed prompts and videos to obtain descriptions and detection results.

A.3.2. LAVAD [12]

LAVAD leverages CLIP to describe each frame's content and aggregates the descriptions of multiple frames to get the final video description.

A.3.3. Zero-Shot CLIP [8] and Zero-Shot ImageBind [2]

These methods utilize CLIP and ImageBind to project text features and image features into the same space, directly comparing the image features extracted by the pre-trained models with normal text features to obtain anomaly scores.

B. More Visualization

We provide further illustrations of the Phys-AD dataset through various figures, highlighting the unique characteristics and challenges it presents.

B.1. Phys-AD Dataset Overview

Figure A shows examples from the Phys-AD dataset. Many deformable, articulated, and assembled objects appear normal in a static state but reveal anomalies only through physical manipulation.

Figure A. **Phys-AD Dataset.** Many deformable, articulated, and assembled objects appear normal in a static state but reveal anomalies only through physical manipulation.

B.2. Anomaly Cases

Figures B and C present anomaly cases from the Phys-AD dataset. These figures illustrate various anomalies that are challenging to detect due to their subtle visual cues and reliance on physical interactions.

B.3. Video Demonstrations

Figures D and E provide video demonstrations from the Phys-AD dataset, showcasing dynamic interactions that highlight the physical properties and anomalies present in the data.

B.4. Generation of PAEval Labels

Figure F illustrates the generation of PAEval labels, demonstrating how labels for physical properties are generated in the dataset.

B.5. Examples of Video-Language Model Descriptions

Figure G shows examples of how video-language models describe videos from the Phys-AD dataset. Correct summaries are highlighted in green, and incorrect ones in red, emphasizing the challenges VLMs face in understanding

these videos.

C. Implementation Details

C.1. Hyperparameters

The modified hyperparameters are detailed below, while others remain the same as specified in the original works:

1. **MemAE** [3]: Batch size = 32, epochs = 30
2. **MPN** [5]: Epochs = 21
3. **MNAD** [7]: Epochs = 7
4. **S3R** [10]: Batch size = 4, epochs = 1000
5. **MGFN** [1]: Epochs = 500

C.2. Prompts

We employed various prompt designs to guide models, as described below.

C.2.1. LAVAD [12]

- **Score Prompt:** "If you were a law enforcement agency, how would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious activities? Provide the response as a Python list, selecting a single number from `[0, 0.1, 0.2, ..., 1.0]`. Do not include any textual explanation."
- **Description Prompt:** "Summarize the events described in the temporal sequence of the scene. Avoid unnecessary details or elaboration."

C.2.2. Video-LLaVA [4]

- **Score Prompt:** "Evaluate the anomaly level in this video on a scale from 0 to 1, where 0 represents 'Absolutely no abnormalities' and 1 represents 'Absolutely abnormal.' Avoid using the exact score of 0.5. Provide only the score without additional explanation."
- **Description Prompt:** "Describe the events depicted in the video in detail."
- **Explanation Prompt:** "Assume you are an anomaly detection expert. Is the function of the object in the video normal or abnormal? Provide a detailed explanation."

C.2.3. Video-LLaMA [13] & Video-ChatGPT [6]

For these models, prompts consist of `COMMON_QUESTIONS`, followed by `CATEGORY_SPECIFIC_QUESTIONS`, depending on the object class. Scoring, description, and explanation are generated in one step. Below is an example for the *ball* class:

- `COMMON_QUESTIONS:`
  1. "What is the object in the video?"
  2. "What is the normal function of the object in real life?"
  3. "What is the mode of interaction observed in the video?"
  4. "Describe the content of this video, focusing on objects, appearance, and physical interactions."
  5. "As an anomaly detection expert, assess whether the object's function is normal or abnormal. Provide a reasonable explanation."
- `CATEGORY_SPECIFIC_QUESTIONS (for ball class):` "Assume the object in the video is a ball. Under normal conditions, a fully inflated ball resists significant deformation. Rate the anomaly on a scale from 0 to 1, where 0 is 'definitely normal' and 1 is 'definitely abnormal.' Provide only the anomaly score in the format: `{anomalyscore=}` without additional text."

C.2.4. PAEval Prompt

- **System Prompt:** "I am an expert in text comparison. I evaluate the semantic similarity of texts, considering spatiotemporal relationships and event structures. I assign a similarity score between 0 and 1, where higher scores indicate greater similarity."
- **User Prompt:** "Given the input text: `{text}`, compare it to entries in the label text library: `{labels}`. Assign a similarity score and output only the highest score as the result."

D. Additional Experimental Results

D.1. Result Overview

Tables A and B summarize the Average Precision (AP) and Accuracy (ACC) of various methods on the Phys-AD dataset across 22 categories. These results cover three methodological paradigms: unsupervised, weakly supervised, and video understanding approaches. It is important to note that both AP and ACC metrics can be influenced by the ratio of positive and negative samples, making these metrics indicative rather than absolute. For ACC, a decision threshold of 0.5 is used by default.

D.2. Observations and Insights

D.2.1. Performance Trends Across Paradigms

- **Unsupervised Methods:** These approaches, such as MNAD.r, achieve competitive results in simpler scenarios, with an average AP of 0.797. However, they often struggle with categories that exhibit complex temporal dynamics or physical interactions.
- **Weakly Supervised Methods:** Methods like S3R and MGFN outperform unsupervised approaches, benefiting from limited supervision. They show consistent improvements in categories requiring higher precision.
- **Video Understanding Models:** Advanced models such as Video-LLaMA and Video-LLaVA demonstrate superior performance by leveraging contextual and semantic reasoning. This is evident in challenging categories such as 'Car' and 'Gear,' where contextual understanding plays a key role.

D.2.2. Category-Specific Insights

- **High Variability:** Categories like 'Sticky Roller' and 'Servo' exhibit significant performance variability across methods, highlighting the challenges of modeling subtle interactions and anomalies.
- **Limitations in Specific Categories:** Categories like 'Rubber Band' and 'USB' present low AP and ACC across all methods, reflecting the difficulty of detecting low-contrast anomalies or deformations.
- **Strengths in Contextual Modeling:** In categories like 'Ball' and 'Magnet,' video understanding models excel, showcasing the advantage of integrating physical reasoning and contextual cues.

D.2.3. Challenges with Balanced Metrics

- The ACC metric is highly sensitive to class imbalance, particularly in categories like 'Hinge' and 'Caster Wheel,' where unsupervised methods often underperform due to skewed distributions.

D.2.4. General Observations

- **Incorporation of Domain Knowledge:** Models incorporating domain-specific knowledge, such as Video-LLaMA and LAVAD, perform significantly better in categories like 'Button' and 'Clip.'
- **Plateaus in Weakly Supervised Performance:** While effective, weakly supervised methods may reach performance ceilings, suggesting the need for more advanced hybrid or fully supervised approaches.

E. Potential Negative Social Impacts

Our dataset was collected with permission from the factory, ensuring compliance with ethical standards. Therefore, we anticipate no negative social impacts arising from this work.

References

[1] 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. 1, 3, 5

[2] 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. 1, 5

[3] 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. 1, 3, 5

[4] 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. 1, 3, 5

[5] 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. 1, 3, 5

[6] 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. 1, 3, 5

[7] 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. 1, 3, 5

[8] 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. 1, 5

[9] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1, 5

[10] 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. 1, 3, 5

[11] 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. 1, 5

[12] 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. 1, 3, 5

[13] 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. 1, 3, 5

Table A. **Video-level AP (↑) result of 22 categories on Phys-AD dataset.** We include Unsupervised, Weakly-supervised and Video-understanding methods.'ZS ImgB','V-ChatGPT','V-LLaMA','V-LLaVA' denote ZS ImageBind,Video-ChatGPT,Video-LLaMA and Video-LLaVA.

Category.	**Unsupervised**					**Weakly-supervised**			**Video-understanding**					
	MPN [5]	MemAE [3]	MNAD.p [7]	MNAD.r [7]	SVM [9]	VADClip [11]	S3R [10]	MGFN [1]	LAVAD [12]	ZS Clip [8]	ZS ImgB [2]	V-ChatGPT [6]	V-LLaMA [13]	V-LLaVA [4]
Car	0.628	0.770	0.762	0.981	0.784	0.787	0.816	0.797	0.773	0.750	0.750	0.751	0.876	0.759
Fan	0.916	0.371	0.933	0.763	0.750	0.800	0.823	0.796	0.757	0.750	0.750	0.763	0.861	0.795
Rolling Bearing	0.522	0.320	0.479	0.812	0.882	0.551	0.619	0.648	0.499	0.500	0.500	0.429	0.741	0.500
Spherical Bearing	0.332	0.382	0.929	0.787	0.588	0.500	0.647	0.596	0.488	0.500	0.500	0.479	0.693	0.500
Servo	0.745	0.795	0.992	0.957	0.750	0.759	0.803	0.808	0.756	0.750	0.750	0.753	0.857	0.745
Clip	0.732	0.619	0.785	0.596	0.667	0.631	0.722	0.720	0.701	0.693	0.667	0.761	0.820	0.649
USB	0.357	0.742	0.646	0.946	0.500	0.530	0.567	0.548	0.512	0.500	0.500	0.535	0.765	0.500
Hinge	0.924	0.965	0.887	0.967	0.750	0.868	0.789	0.834	0.794	0.758	0.750	0.750	0.894	0.750
Sticky Roller	0.989	0.989	0.698	0.975	0.667	0.686	0.883	0.829	0.553	0.625	0.667	0.645	0.834	0.656
Caster Wheel	0.730	0.815	0.644	0.823	0.750	0.797	0.876	0.901	0.820	0.735	0.750	0.732	0.891	0.750
Screw	0.522	0.763	0.750	0.685	0.667	0.667	0.769	0.720	0.763	0.680	0.667	0.658	0.826	0.690
Lock	0.741	0.683	0.623	0.795	0.789	0.662	0.704	0.767	0.586	0.667	0.667	0.613	0.831	0.667
Gear	0.839	0.826	0.874	0.865	0.800	0.807	0.829	0.818	0.603	0.800	0.800	0.816	0.903	0.805
Clock	0.572	0.751	0.614	0.711	0.670	0.670	0.708	0.698	0.684	0.667	0.670	0.667	0.842	0.670
Slide	0.806	0.991	0.978	0.942	0.667	0.817	0.844	0.864	0.772	0.800	0.800	0.817	0.906	0.726
Zipper	0.898	0.669	0.896	0.674	0.667	0.669	0.777	0.757	0.712	0.667	0.667	0.715	0.832	0.667
Button	0.966	0.726	0.903	0.853	0.800	0.845	0.818	0.830	0.778	0.842	0.800	0.804	0.905	0.763
Liquid	0.564	0.890	0.860	0.927	0.667	0.686	0.726	0.900	0.712	0.784	0.667	0.622	0.800	0.579
Rubber Band	0.411	0.410	0.433	0.394	0.536	0.491	0.670	0.631	0.499	0.500	0.500	0.509	0.751	0.478
Ball	0.716	0.661	0.842	0.826	0.667	0.667	0.809	0.727	0.739	0.667	0.667	0.699	0.839	0.682
Magnet	0.746	0.754	0.802	0.603	0.667	0.667	0.737	0.841	0.673	0.667	0.667	0.780	0.860	0.626
Toothpaste	0.657	0.899	0.653	0.644	0.500	0.500	0.682	0.746	0.569	0.500	0.500	0.464	0.763	0.484
Average	0.703	0.735	0.772	0.797	0.690	0.684	0.755	0.763	0.681	0.673	0.666	0.671	0.831	0.656

Table B. **Video-level ACC (↑) result of 22 categories on Phys-AD dataset.** We include Unsupervised, Weakly-supervised and Video-understanding methods.'ZS ImgB','V-ChatGPT','V-LLaMA','V-LLaVA' denote ZS ImageBind,Video-ChatGPT,Video-LLaMA and Video-LLaVA.

Category.	**Unsupervised**					**Weakly-supervised**			**Video-understanding**					
	MPN [5]	MemAE [3]	MNAD.p [7]	MNAD.r [7]	SVM [9]	VADClip [11]	S3R [10]	MGFN [1]	LAVAD [12]	ZS Clip [8]	ZS ImgB [2]	V-ChatGPT [6]	V-LLaMA [13]	V-LLaVA [4]
Car	0.703	0.750	0.755	0.753	0.793	0.402	0.262	0.250	0.332	0.750	0.750	0.687	0.504	0.447
Fan	0.825	0.750	0.769	0.750	0.750	0.785	0.302	0.250	0.311	0.750	0.750	0.735	0.446	0.667
Rolling Bearing	0.450	0.500	0.433	0.500	0.933	0.589	0.558	0.500	0.500	0.500	0.500	0.300	0.446	0.500
Spherical Bearing	0.417	0.500	0.883	0.567	0.650	0.682	0.538	0.500	0.500	0.500	0.500	0.450	0.346	0.500
Servo	0.742	0.795	0.769	0.750	0.750	0.277	0.344	0.250	0.250	0.750	0.750	0.675	0.420	0.237
Clip	0.667	0.619	0.667	0.667	0.667	0.462	0.330	0.330	0.333	0.693	0.667	0.648	0.467	0.531
USB	0.467	0.500	0.508	0.504	0.500	0.530	0.513	0.500	0.496	0.500	0.500	0.563	0.518	0.500
Hinge	0.750	0.750	0.750	0.750	0.750	0.686	0.438	0.229	0.283	0.758	0.750	0.250	0.564	0.250
Sticky Roller	0.667	0.667	0.667	0.667	0.667	0.694	0.667	0.333	0.333	0.625	0.667	0.556	0.518	0.333
Caster Wheel	0.750	0.750	0.533	0.733	0.750	0.765	0.346	0.212	0.250	0.735	0.750	0.650	0.575	0.250
Screw	0.667	0.667	0.667	0.667	0.667	0.667	0.424	0.333	0.733	0.680	0.667	0.444	0.496	0.578
Lock	0.633	0.667	0.667	0.667	0.822	0.304	0.333	0.339	0.328	0.667	0.667	0.244	0.470	0.333
Gear	0.798	0.800	0.738	0.809	0.800	0.244	0.230	0.200	0.209	0.800	0.800	0.780	0.520	0.793
Clock	0.670	0.670	0.661	0.670	0.670	0.670	0.439	0.330	0.344	0.667	0.670	0.511	0.521	0.330
Slide	0.800	0.800	0.800	0.800	0.667	0.291	0.410	0.194	0.193	0.800	0.800	0.730	0.494	0.267
Zipper	0.661	0.667	0.717	0.667	0.667	0.526	0.345	0.333	0.344	0.667	0.667	0.555	0.508	0.334
Button	0.800	0.800	0.797	0.800	0.800	0.596	0.264	0.197	0.197	0.842	0.800	0.470	0.499	0.317
Liquid	0.667	0.667	0.667	0.667	0.667	0.639	0.515	0.333	0.333	0.784	0.667	0.477	0.363	0.289
Rubber Band	0.500	0.500	0.500	0.500	0.567	0.482	0.519	0.500	0.500	0.500	0.500	0.517	0.515	0.450
Ball	0.667	0.577	0.667	0.667	0.667	0.667	0.374	0.333	0.333	0.667	0.667	0.570	0.523	0.644
Magnet	0.667	0.667	0.689	0.667	0.667	0.548	0.485	0.333	0.333	0.667	0.667	0.600	0.580	0.489
Toothpaste	0.500	0.500	0.500	0.500	0.500	0.500	0.545	0.500	0.500	0.500	0.500	0.400	0.559	0.467
Average	0.658	0.735	0.673	0.669	0.699	0.543	0.417	0.331	0.361	0.673	0.666	0.537	0.493	0.432

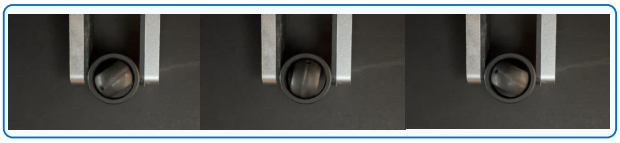
Normal: The rail slides smoothly without any interruptions.

Normal: The rail slides smoothly and well-structured.

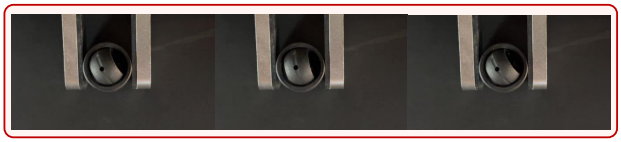
Abnormal: A ball disengages from the slide rail and falls.

(a) Intermittent Anomalies of Slides

Normal: The spherical bearing rotates smoothly as intended.

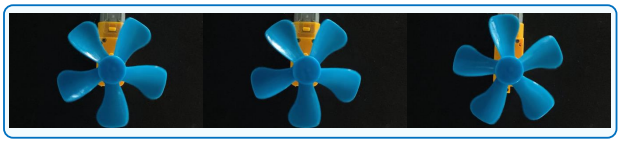


Abnormal: The spherical bearing unexpectedly be stuck and halts its movement.



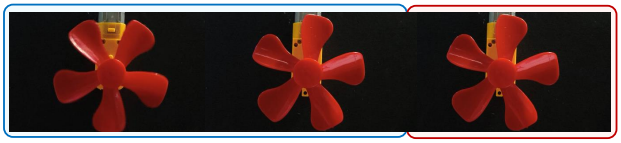
(b) Persistent Anomalies of Spherical Bearing

Normal: The fan rotates steadily.



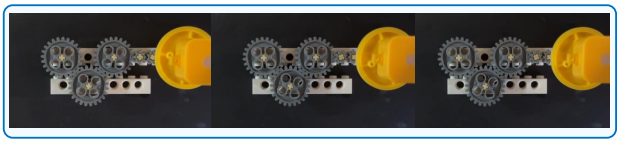
Normal: The fan spins rapidly at a steady speed .

Abnormal: The fan stuck and can no longer rotate.



(c) Intermittent Anomalies of Fan

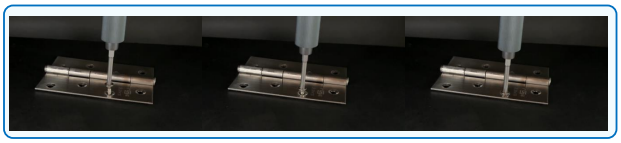
Normal: The gear turns steadily in response to the motor's drive.



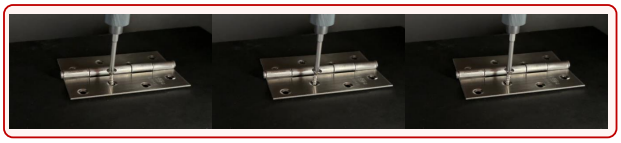
Abnormal: One of the gear remains stationary and not mesh.

(d) Persistent Anomalies of Gear

Normal: The screw tightens gradually.

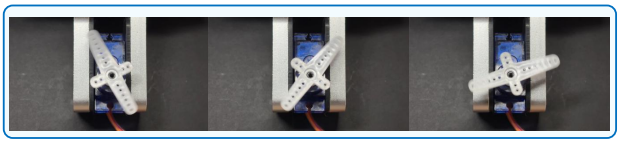


Abnormal: The screw remains unable to fit into the hole.



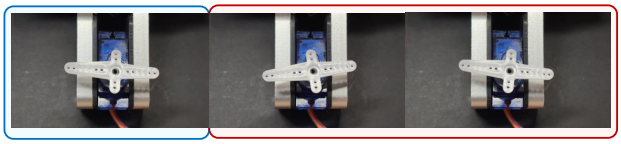
(e) Persistent Anomalies of Screw

Normal: The Servo rotates smoothly.



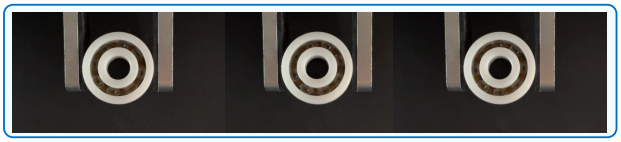
Normal: The Servo rotates smoothly at the beginning.

Abnormal: The swing angle is restricted.

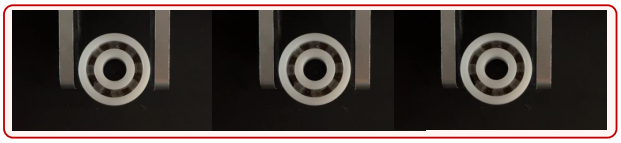


(f) Intermittent Anomalies of Servo

Normal: The rolling bear spins, causing the inner ball to rotate.



Abnormal: The internal ball does not rotate.

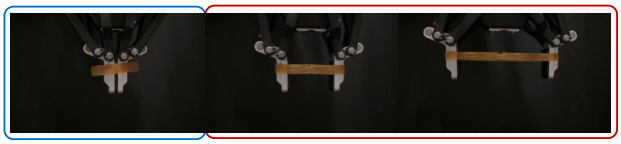


(g) Persistent Anomalies of Rolling Bear

Normal: The rubber band stretches and lengthens.

Normal: The rubber band stretches .

Abnormal: Small cracks begin to appear on the rubber band.

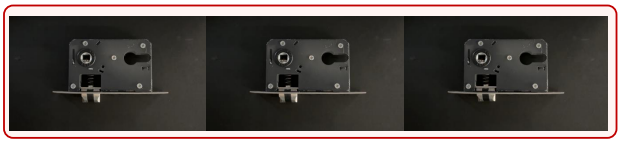


(h) Intermittent Anomalies of Rubber Band

Figure B. **Anomaly Cases of Phys-AD Dataset (1/2).**

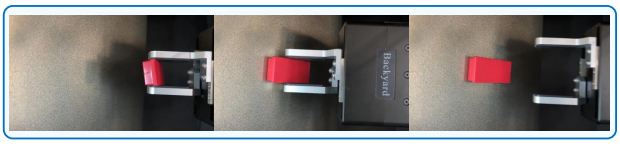
Normal: The lock's keyhole rotates, the bolt extends and retracts.

Abnormal: The lock's keyhole and bolt are stuck.



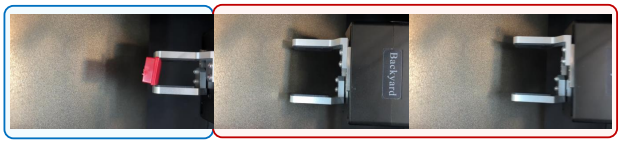
(k) Persistent Anomalies of Lock

Normal: The magnet attaches to the metal plate.



Normal: The magnet moves toward to a metal plate.

Abnormal: The magnet unexpectedly falls off.



(l) Intermittent Anomalies of Magnet

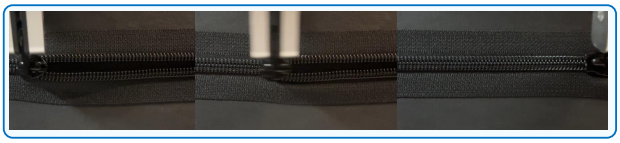
Normal: The toothpaste is pressed and deformed.

Normal: The toothpaste is squeezed.

Abnormal: The toothpaste begins to leak out.

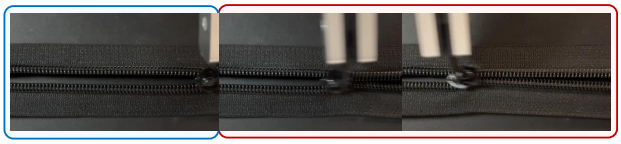
(m) Intermittent Anomalies of Toothpaste

Normal: The zipper to open and close smoothly.



Normal: The zipper glides smoothly at the beginning.

Abnormal: The zipper remains unable to close fully.



(n) Intermittent Anomalies of Zipper

Normal: The hinge rotates back and forth normally .

Normal: The hinge starts to rotate.

Abnormal: The hinge shaft becomes loose and falls away.

(o) Intermittent Anomalies of Hinge

Normal: The liquid inside the bottle is clear and does not leak.

Normal: The bottle is shaken resulting in a change in the water level.

Abnormal: The liquid leaks out.

(p) Intermittent Anomalies of Liquid

Normal: the car's wheels maintain a steady rotation.

Abnormal: The front wheels turn freely but the back wheels do not rotate.

(q) Persistent Anomalies of Car

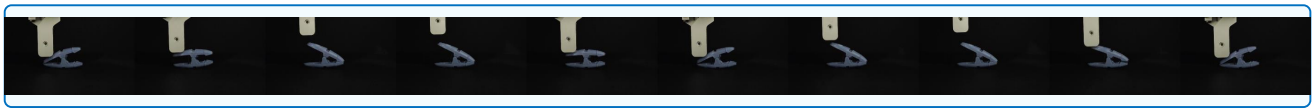
Normal: The caster wheel 's every aixs rotates as expected.

Abnormal: The axis of caster wheel appears to be stuck and unable to rotate.

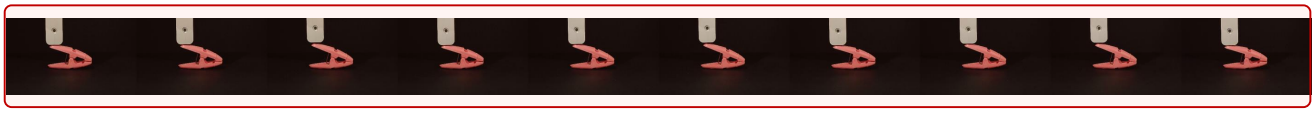
(r) Intermittent Anomalies of Caster Wheel

Figure C. **Anomaly Cases of Phys-AD Dataset (2/2).**

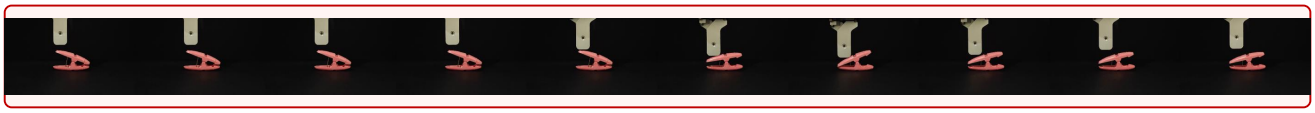
Normal: The video shows a robotic arm interacting with a clip. The arm exerts pressure on the clip, causing it to open. When the robotic arm releases the pressure, the clip automatically returns to its original closed position.



Abnormal: In the video, clips and a robotic arm are featured. The robotic arm exerts pressure on the clips, but the clips remain closed and do not open, which is consistent with their original state.

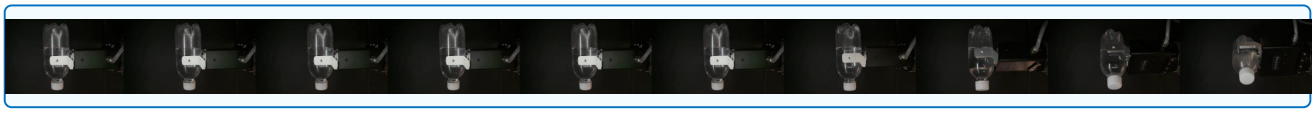


Abnormal: The video shows a clip and a robotic arm working together. As the robotic arm presses down on the clip, it opens at a specific angle. When the robotic arm is removed, the clip stays open and does not revert to its initial, closed position.

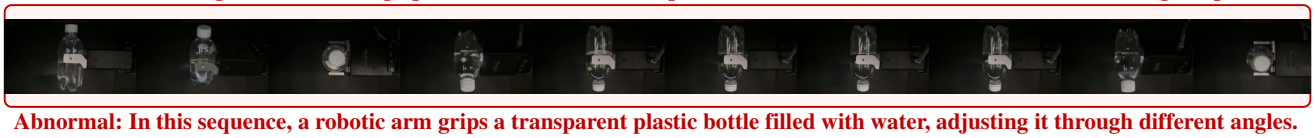


(a) Video Demo of Clip

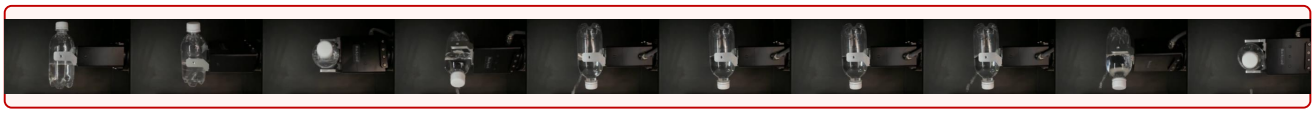
Normal: The video shows a robotic arm that engages with a transparent, water-filled bottle through multiple actions: placing the bottle, inverting it to pour water into a clear glass, and shaking it. The water within the bottle stays clear and devoid of any foreign matter throughout the procedure.



Abnormal: In this video, a robotic arm handles a clear bottle containing water, moving it through various motions like positioning, inverting, and shaking. Upon turning the bottle upside down, water pours out, exposing a small light and a white floating object within. While shaking the bottle, foreign particles are seen in the liquid, and at the end, the bottle is set back to its original position.

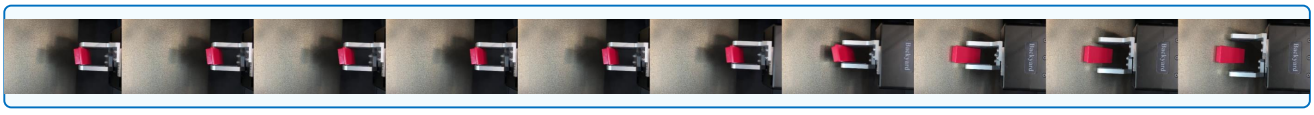


Abnormal: In this sequence, a robotic arm grips a transparent plastic bottle filled with water, adjusting it through different angles. Initially, the arm flips the bottle upside down, causing water to pour out. It then returns the bottle to an upright position, resulting in a drop in the water level. Later, the arm shakes the bottle, leading to further leakage.



(b) Video Demo of Liquid

Normal: In the video, a robotic arm is seen interacting with a magnet. The robotic arm securely holds the magnet and moves it closer to a metal plate. Upon releasing the magnet, it is immediately drawn to and attaches to the metal plate.

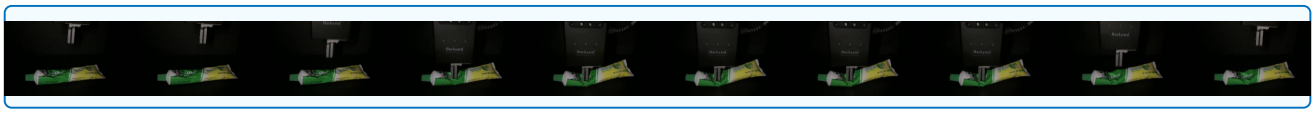


Abnormal: In the video, a robotic arm interacts with a magnet, gripping it and moving it close to a metal plate. When the robotic arm releases its hold, the magnet doesn't stick but instead falls off immediately.

(c) Video Demo of Magnet

Figure D. **Video Demo of the Phys-AD Dataset (1/2).**

Normal: In the video, a robotic arm is seen working with a toothpaste tube, pressing and kneading it repeatedly. With each motion of the robotic arm, the toothpaste deforms accordingly.

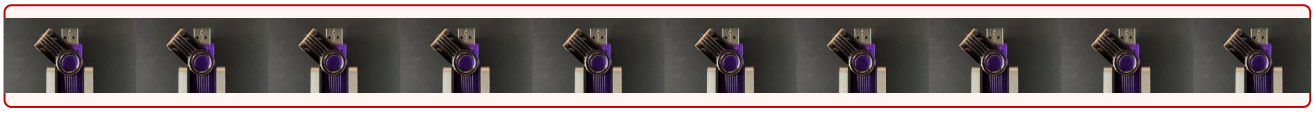


Abnormal: The video presents a robotic arm gripping a tube of toothpaste, pressing and squeezing it. As the robotic arm continues these actions, the toothpaste becomes deformed and begins to leak with each squeeze.

(a) Video Demo of Toothpaste

Normal: The video presents a USB drive whose cover is powered by a motor. With each rotation, the motor enables the cover to swing back and forth smoothly.

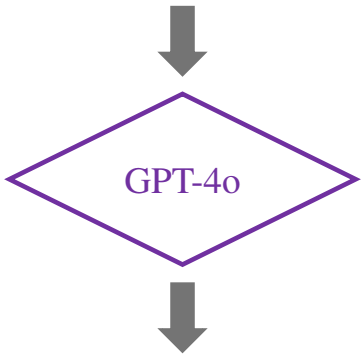
Abnormal: The video presents a USB drive with a motor powering its cover to swing. Although it begins moving back and forth, the cover becomes lodged at a specific angle, unable to complete the swing.



(b) Video Demo of USB

Figure E. **Video Demo of the Phys-AD Dataset (2/2).**

In the video, a robotic arm interacts with a clear bottle filled with water, performing several actions including positioning the bottle, turning it upside down to pour the water into a clear glass, and shaking it. Throughout the entire process, the water inside the bottle remains clear and free of any foreign particles.



1.In the video, a robotic arm interacts with a clear bottle filled with water, performing several actions including positioning the bottle, turning it upside down to pour the water into a clear glass, and shaking it. Throughout the entire process, the water inside the bottle remains clear and free of any foreign particles.

2.The video features a robotic arm that engages with a transparent, water-filled bottle through multiple actions: placing the bottle, inverting it to pour water into a clear glass, and shaking it. The water within the bottle stays clear and devoid of any foreign matter throughout the procedure.

3.A robotic arm is shown interacting with a clear bottle containing water, executing various actions such as positioning it, flipping it upside down to pour water into a transparent glass, and shaking the bottle. Throughout these actions, the water remains clear with no foreign particles present.

4.In the video, a robotic arm manipulates a transparent bottle filled with water in several ways, including correctly positioning it, turning it upside down to pour water into a clear glass, and shaking it. The water inside remains clear and free from any impurities during the entire process.

5.The video depicts a robotic arm interacting with a clear, water-filled bottle, engaging in multiple actions like placing the bottle, inverting it to pour water into a transparent glass, and shaking it. Throughout the sequence, the water in the bottle stays clear and contains no foreign particles.

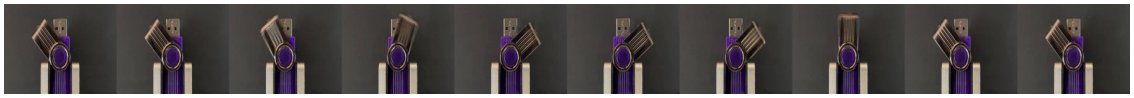
Figure F. **Generation of PAEval Labels (Labels for Liquid in the Graph).**

VLM

In the video, there is a plastic bottle that is the main object of focus. The bottle is white and has a distinct shape, with two parallel ridges on its body. The bottle is also translucent, allowing us to see the inside of the container.

GT

In the video, a robotic arm interacts with a clear bottle filled with water, performing several actions including positioning the bottle, turning it upside down to pour the water into a clear glass, and shaking it. Throughout the entire process, the water inside the bottle remains clear and free of any foreign particles.



VLM

In the video, a purple USB flash drive **is being inserted into a USB port on a computer**. The flash drive is being inserted into the port, and once it is fully inserted, it is seen sitting in the port. The video does not show any other objects or actions, and **the focus is solely on the USB flash drive** being inserted into the port.

GT

In the video, a USB drive is featured as the primary object. The USB cover, powered by a motor, swings back and forth with a smooth and consistent motion.

VLM

The video features **a train** on a dark background. The train has a smooth, metallic appearance with round wheels and a long, **flat body**. The train moves and rotates on its wheels in a slow, steady motion.

GT

In the video, a door lock is shown in action. A motor drives the rotation of the keyhole, which in turn causes the door bolt to extend outward and retract back with each complete rotation.

Figure G. **Examples of anomaly descriptions generated by Video-LLaMA.** Green text indicates correct summaries, while red text indicates incorrect ones.