

O-VAD: Industrial Video Anomaly Detection through Object-Centric Tracking and Reasoning

Mei Yuan¹, Qi Long¹, Qifeng Wu¹, Zhenyang Li², Yizhou Zhao¹, Lei Wang³,
Yang Liu¹, and Min Xu^{1*}

¹ Carnegie Mellon University

² University of Alabama at Birmingham

³ Griffith University

meiyuan@andrew.cmu.edu, mxu1@cs.cmu.edu

<https://o-vad.github.io/>



Fig. 1: Why object state evolution matters for industrial anomaly detection. The toothpaste tube undergoes pressing and deformation before a subtle leakage emerges. Traditional methods predict binary classifications. VLM prompting fails to detect it due to the lack of evidence. O-VAD tracks object-wise state changes and produces an open-ended anomaly report with in-depth analysis.

Abstract. Industrial Video Anomaly Detection (IVAD) aims to identify anomalous objects and events in an industrial process, which is crucial for modern manufacturing and quality control systems. Existing VLM-based anomaly reasoning methods are capable of detecting open-ended anomalies in general domains. However, their performance declines in industrial settings characterized by intricate object transformations, strict physics, and procedural constraints. To tackle the complexity of such interaction-intensive detection, we introduce a training-free agentic framework **O-VAD** for anomaly detection free of domain-specific knowledge, emphasizing object state evolution like humans inspectors. It is designed to

* Corresponding author.

track spatial-temporal dynamics and underlying transformations of detected objects over time, and then reason over the object-wise temporal state trajectories to identify abnormal objects in grounded frames. Our method overcomes limitations of prior approaches that rely on retraining on normal clips or injecting domain knowledge as context for test-time inference. Extensive experiments on three IVAD datasets demonstrate that our method outperforms frontier VLMs, agentic frameworks, and traditional VAD methods fine-tuned on the respective datasets, while providing interpretable reports over anomaly processes and types.

Keywords: Industrial Anomaly Detection · Agentic Reasoning · Vision-Language Models

1 Introduction

Industrial Video Anomaly Detection (IVAD) aims to identify anomalous objects and events in an industrial video process. It plays a critical role in modern manufacturing and quality control systems, where automated visual inspection can significantly reduce costs and improve production reliability [26]. Industrial processes are extremely challenging and different from general domains, since they are usually characterized by: (i) *complex object transformations*: objects undergo substantial physical and functional state changes through cutting, assembly, heating, pressing, and other manufacturing operations; (ii) *strict physical and procedural constraints*: anomalies manifest as violations of expected spatial configurations, periodic procedures, or physical laws; (iii) *high interpretability requirements*: operators need explicit explanations of detected anomalies for root cause analysis and corrective actions.

Traditional Video Anomaly Detection (VAD) methods primarily adopt two paradigms: reconstruction-based approaches [8, 9, 12, 42–44] reconstruct anomalous samples to their corresponding normal counterparts and calculate the reconstruction error, and embedding-based [6, 11, 15, 17, 32] methods focus on modeling the feature embeddings of normal samples and measure deviations. But they typically follow the “one-class-one-model” learning paradigm, requiring plentiful normal samples for each object class to learn its distribution [10], which makes it impractical for industrial anomaly detection settings and less suitable for dynamic production environments. Given that ***traditional VADs have limited generalization abilities and binary classifications provide no semantic rationale for their decisions***, recent works [10, 14, 18, 24, 25, 45, 48, 49] propose a series of VLM-based anomaly detection methods, featuring semantic anomaly understanding and test-time detection inference.

Although VLMs have unlocked their potential for detecting anomalies augmented with semantic-visual understanding in general domains, there are still two issues that remain up in the air: (1) Their performance in industrial settings declined due to *lack of domain-specific knowledge and fine-grained annotations*. Recent training-free studies [16] provide external knowledge or normal samples in context for test-time inference to alleviate this limitation. However, ***they***

heavily rely on sophisticated prompt design to caption videos, inject knowledge, and then assign anomaly scores by off-the-shelf VLMs. The overly dependence on external context leads to misaligned responses that prioritize contextual plausibility over accuracy [18]. (2) Video/frame level features or descriptions have *weak understanding of localized details within objects over time*, which hinders their reliable and accurate interaction-intensive IVAD reasoning. Some VLM-based methods exploited the spatio-temporal fusion [20,27,34,38,39], and detect anomalies from when and where perspectives [13,24]. However, *they lose focus on object-centric physics and interactions over time*. It is the core when human inspectors detect object-centric state changes throughout industrial processes, especially when objects undergo substantial physical and functional state changes.

To tackle these two issues, we propose **O-VAD**, a training-free agentic framework that detects anomalies by tracking object state evolution over time—free of any domain-specific knowledge, labels, or predefined anomaly taxonomy. (1) For the first issue, rather than injecting external knowledge or relying on complex prompt design, O-VAD enables the VLM to *build its own understanding* through a “ground→track→reason” agentic pipeline. Objects are first detected and segmented via **VLM-grounded masking** with SAM3 [2], then continuously tracked and queried across frames to accumulate structured evidence. A cascaded chain-of-thought (CoT) then derives anomaly judgments from the VLM’s own internalized commonsense and physical knowledge, mirroring the cognitive process of human inspectors. (2) For the second issue, we shift the unit of analysis from holistic frames to individual objects. An **object state tracker** maintains per-object representations and queries the VLM at key temporal transitions to capture fine-grained state changes, *e.g.* deformation, material release, surface alteration. These structured state trajectories ground the downstream reasoning, making detection both *object-centric* and *spatio-temporally aware*, and yielding *open-ended, fine-grained* outputs: anomaly types, grounded abnormal frames, affected objects, and causal analyses.

Result-wise, despite using no labor-intensive annotations, the proposed agentic framework delivers state-of-the-art performance on any level anomaly detection, surpassing frontier VLMs and agentic frameworks on quantitative and qualitative evaluations.

To summarize our contributions:

(1) We present a generalizable data curation pipeline and apply it to three object-centric IVAD datasets (IPAD, Phys-AD and LiquidAD) with fine-grained annotations of object trajectories across frames.

(2) We propose a training-free, three-stage agentic reasoning framework that requires no domain-specific knowledge or fine-tuning, performing object discovery, spatiotemporal tubelet construction with open-ended state change detection, and chain-of-thought anomaly reasoning with visual verification.

(3) Our framework achieves state-of-the-art performance on three IVAD datasets at both video- and frame- levels, outperforming frontier VLMs (Qwen3-VL-32B and GPT-5), two agentic framework for VAD, and fine-tuned traditional

AD methods, while providing interpretable reasoning over anomaly processes and open-ended types.

2 Related Work

Traditional Video Anomaly Detection. Traditional video anomaly detection methods primarily adopt two paradigms: reconstruction-based approaches [8, 9, 12, 42–44] flag anomalies by their elevated reconstruction error, while embedding-based methods [6, 11, 15, 17, 32] model the feature distribution of normal samples and measure deviations from it. Despite strong benchmark performance, both share three limitations. First, they follow a “one-class-one-model” paradigm [10], requiring abundant normal samples and a separate model per object class, and must be re-trained for every unseen domain or anomaly type [22], making them impractical for novel categories and dynamic production lines. Second, they are largely black-box, emitting an anomaly score without semantic rationale [39], which has motivated growing interest in explainable, language-grounded anomaly understanding [22, 46]. Third, they compress entire frames into holistic features, neglecting region-level cues and over-relying on dominant background context [21]. Our O-VAD addresses these limitations with a training-free, object-centric framework that requires no per-class training, produces semantic explanations for its decisions, and reasons over region-level evidence.

Multimodal Video Anomaly Detection Recent works [10, 14, 18, 24, 25, 45, 48, 49] revisit anomaly detection with VLM-based methods, valued for their generalizability and explainability, especially in the zero/few-shot setting. These approaches rely on sophisticated prompt design that captions videos and assigns anomaly scores via off-the-shelf VLMs. To strengthen spatio-temporal awareness, some methods couple this with auxiliary temporal modeling or prompting [20, 27, 34, 38, 39]; for instance, a unified zero-shot framework [24] chains temporal detection, spatial localization, and textual explanation through a test-time reasoning process over foundation models, requiring no additional training. Yet these methods largely overlook that superficial temporal caption sequences lose focus on the object-centric dynamics and interactions over time—precisely the cues human inspectors rely on to detect anomalous objects and events throughout industrial processes. In contrast, our O-VAD explicitly models object-centric spatio-temporal evidence while retaining the training-free and explainable advantages of VLM-based detection.

3 Method

Overview. Given an industrial process video $\mathcal{V} = \{I_t\}_{t=1}^T$, O-VAD produces a structured anomaly report for each detected anomaly, it outputs the anomaly type, severity, affected object, temporal localization, and a natural-language causal explanation—all without any domain-specific knowledge, predefined taxonomy, or training data. To achieve this, O-VAD adopts a three-stage agentic

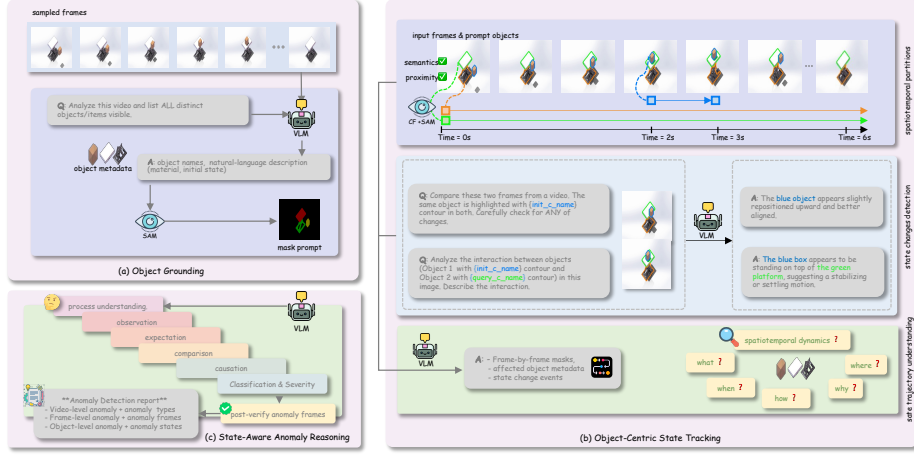


Fig. 2: Overview of O-VAD. (a) Stage 1 samples multiple frames and queries a VLM to discover all objects, then segments them with SAM to produce initial masks and metadata. (b) Stage 2 constructs spatiotemporal tubelets via CropFormer [29] and SAM2 [31], recovers missing tracks through semantic and proximity priors, and query the VLM for open-ended state change detection and inter-object interaction analysis. (c) Stage 3 feeds the accumulated object metadata and state change events into a multi-step chain-of-thought reasoning chain that distinguishes process actions from failure outcomes, followed by visual verification of candidate anomalies.

pipeline (Fig. 2). Stage 1 (§3.2) discovers and segments all task-relevant objects via VLM-grounded masking with SAM [2], rather than manual annotation. Stage 2 (§3.3) tracks each object through potential state trajectories via spatiotemporal partitioning, querying the VLM at temporal transitions to detect fine-grained state changes and produce per-object state trajectories. Stage 3 (§3.4) reasons over the accumulated evidence through a multi-step chain-of-thought that separates abnormal stage changes from expected process outcomes, yielding anomaly judgments grounded in the VLM’s own commonsense and physical reasoning.

3.1 Problem Formulation

Our O-VAD decomposes the problem into an object-centric formulation. We first detect K objects $\mathcal{O} = \{o_k\}_{k=1}^K$ with masks $\{m_k^{(t)}\}$ via VLM-grounded segmentation, then track each object’s state trajectory $\mathcal{S}_k = \{(m_k^{(t)}, \phi_k^{(t)})\}_{t=1}^T$, where $m_k^{(t)}$ is the mask of object k at frame t and $\phi_k^{(t)}$ is a VLM-generated natural-language description of its physical state (*e.g.*, name, material, surface condition, and detected state changes). Object-level anomaly reasoning is then performed as:

$$\hat{y}_{t,k} = \theta_{\text{VLM}}(p_{\text{CoT}} \oplus \mathcal{S}_k \oplus c), \quad \hat{y}_t^{\text{frm}} = \max_k \hat{y}_{t,k}, \quad \hat{y}^{\text{vid}} = \max_t \hat{y}_t^{\text{frm}}, \quad (1)$$

where $\hat{y}_{t,k} \in [0, 1]$ is the anomaly confidence for object k at frame t , c is the optional video caption, p_{CoT} is a cascaded chain-of-thought prompt that guides the VLM through cognitive anomaly reasoning, and $\oplus$ denotes concatenation of the prompt context. The frame- and video-level scores are obtained by max-pooling over objects and frames, respectively. By conditioning on per-object state trajectories $\mathcal{S}_k$ rather than raw frames alone, O-VAD produces fine-grained, open-ended anomaly predictions with grounded frames, affected objects, and causal explanations; the discrete anomaly set $\mathcal{A}$ used for reporting is obtained from these scores after the reasoning and verification of Stage 3 (§3.4).

3.2 Stage 1: Automated Object Grounding

Rather than relying on a single reference frame, we sample a set of candidate frames $\{I_{t_j}\}_{j=1}^{N_f}$ spanning the video so that objects initially occluded or absent can still be discovered. For each candidate frame I_{t_j} , a VLM generates a structured object inventory:

$$\mathcal{O}_{t_j} = \phi_{\text{VLM}}(I_{t_j}) = \{(n_k, d_k, b_k)\}_{k=1}^{N_{t_j}}, \quad (2)$$

where n_k is the object name, d_k a natural-language description (material, initial state), and b_k a spatial cue. Inventories across frames are merged into a deduplicated set $\mathcal{O} = \bigcup_j \mathcal{O}_{t_j}$. Each object $o_k \in \mathcal{O}$ is then segmented in a reference frame I_1 using SAM3 [2]:

$$m_k^{(1)} = \text{SAM}(I_1, o_k), \quad \mathcal{M}_1 = \{m_k^{(1)}\}_{k=1}^{|\mathcal{O}|}. \quad (3)$$

The output of Stage 1, including initial masks $\mathcal{M}_1$ together with object meta-data, grounds all subsequent tracking and reasoning in concrete, per-object evidence.

3.3 Stage 2: Object-Centric State Tracking

The second stage tracks each object across all frames, detecting and describing state changes even when objects undergo significant transformations. We build upon TubeletGraph [35] for tracking through transformations, and extend it with two novel components: object interaction analysis and open-ended state change annotation.

Spatiotemporal Partitioning and Tubelet Construction. We construct a spatiotemporal partition $\mathcal{P}$ of the video that associates every pixel region across all frames with a tracked entity. An entity segmentation model (CropFormer [29]) produces per-frame spatial partitions $\mathcal{E}_t = \text{CF}(I_t)$, and each entity $e_1^i \in \mathcal{E}_1$ (including the initial object masks $\mathcal{M}_1$) is tracked forward via SAM2 [31] to form tubelets:

$$\mathcal{P}_{\text{init}} = \{P_i\}_{i=1}^{|\mathcal{E}_1|}, \quad P_i = \{e_t^i\}_{t=1}^T. \quad (4)$$

New tubelets are incrementally added whenever track-less regions emerge. For each entity $\hat{e}_t^j \in \mathcal{E}_t$ at frame $t > 1$, a new tubelet is initiated if less than τ_{coverage} of its area is covered by existing tubelets.

State Tracking. To recover missing object tracks after state transformations (e.g., an object breaking apart or releasing material), candidate tracks that emerge after the initial frame are evaluated using spatial proximity and semantic consistency priors.

Spatial Proximity. For a candidate track $C = \{c_s, c_{s+1}, \dots, c_T\}$ beginning at frame s and the prompt object track $P = \{p_1, p_2, \dots, p_T\}$, the spatial proximity score is:

$$S_{\text{prox}}(C, P) = \max_{j \in \{1, 2, 3\}} \frac{|c_s \cap m_s^j|}{|c_s|}, \quad (5)$$

where $\{m_s^j\}_{j=1}^3$ are the three candidate masks from SAM2’s multi-mask output at frame s . A candidate is considered proximal if $S_{\text{prox}}(C, P) > \tau_{\text{prox}}$.

Semantic Consistency. For a mask M and frame I , we compute the masked CLIP [30] feature $f(M, I) = \text{Pool}(\text{CLIP}(I), M)$ via mask-pooling. The semantic similarity is:

$$S_{\text{sem}}(C, P) = \max_{\substack{i \in \{1, \dots, s-1\}, \\ j \in \{s, \dots, T\}}} f(p_i, I_i) \cdot f(c_j, I_j)^\top. \quad (6)$$

The final set of valid continuation tracks is:

$$\mathcal{V} = \{C \in \mathcal{P} \mid C \text{ begins at } t > 0, S_{\text{prox}}(C, P) > \tau_{\text{prox}}, S_{\text{sem}}(C, P) > \tau_{\text{sem}}\}. \quad (7)$$

The complete tracking result $\mathcal{T} = P \cup \mathcal{V}$ captures the object through transformations.

State Change Detection and Understanding. For object-aware frame pairs, the VLM is prompted with masked visualizations of both frames and asked to identify state changes through visual understanding. Critically, we adopt an *open-ended labeling scheme*: both the change type (e.g., compression deformation, material release, rotational resistance) and the change cause (e.g., object leakage, environmental lighting) are generated as free-form natural language by the VLM. This enables the system to describe novel phenomena for generalizing across diverse industrial processes. Each detected state change event e is represented as a tuple:

$$e = (t_{\text{start}}, t_{\text{end}}, \text{type}, \text{cause}, \text{desc}, \text{sev}, k), \quad (8)$$

where t_{start} and t_{end} are the bounding frame indices, *type* is the free-form change type, *cause* is the causal label, *desc* is a natural-language description, *sev* $\in \{\text{none}, \text{slight}, \text{moderate}, \text{severe}\}$ is the severity, and k is the affected object index.

3.4 Stage 3: State-Aware Anomaly Reasoning

We formulate *cognitive anomaly reasoning* as a cascaded chain-of-thought (CoT) task over the accumulated object states. The final stage synthesizes the tracking metadata, state change events, and visual evidence into an anomaly detection

decision via multi-step reasoning. The key design principle is open-ended classification: the VLM is not constrained to a fixed taxonomy but reasons freely from its domain knowledge, grounded in the tracking evidence. Finally, we have both the anomaly set and the complete reasoning trace for full interpretability.

Cognitive Anomaly Reasoning. The core anomaly reasoning proceeds through a cascaded chain-of-thought (CoT) that mirrors how a human quality inspector diagnoses failures. The VLM receives, in a single call, the object metadata, filtered state changes, inferred task context, video caption, and sampled frames, and is instructed to reason through sequential steps with confidence scores.

1. *Process Understanding.* The VLM identifies the industrial process and its purpose, then triages each state change into expected process, potential anomaly, and normal noise.
2. *Observation.* The VLM cites specific object IDs, frame ranges, change types, and severities from the tracking data, focusing on events triaged as potential anomalies.
3. *Expectation.* The VLM articulates the pass/fail criteria for the identified test, distinguishing expected mechanical responses from failure-indicating outcomes.
4. *Comparison.* For each candidate anomaly, the VLM determines whether it represents a genuine failure outcome or a mechanical response that was mis-categorized.
5. *Causation.* The VLM reasons about root causes without being constrained to predefined categories.
6. *Classification & Severity.* A free-form anomaly type is assigned and each confirmed anomaly receives a severity rating based on its impact on task completion, safety implications, and reversibility.

The reasoning chain outputs a set of candidate anomalies $\{a\}$, where each a carries a free-form anomaly type, an affected object index, a frame range, a severity level, and an initial confidence $c_{\text{orig}}(a) \in [0, 1]$ produced by the VLM. These candidates are then refined by the verification step below to form the final anomaly set $\mathcal{A}$.

Post Visual Verification. To suppress false positives, each candidate anomaly a is gated by its initial confidence $c_{\text{orig}}(a)$: anomalies with $c_{\text{orig}}(a) > \tau_{\text{hi}}$ bypass verification (they are sufficiently certain), those with $c_{\text{orig}}(a) < \tau_{\text{lo}}$ are discarded outright, and only the remaining candidates are verified against their evidence frames:

$$(\text{verified}, c_{\text{ver}}) = \phi_{\text{VLM}}(\{I_t\}_{t \in \text{evidence}}, c, a). \quad (9)$$

The verification step explicitly checks whether the claimed anomaly is actually a normal process behavior visible in the caption c or frames. The final confidence is then computed via a multiplicative gating scheme:

$$c_{\text{final}}(a) = \begin{cases} c_{\text{orig}}(a) \cdot c_{\text{ver}} & \text{if } \text{verified} = \text{true}, \\ c_{\text{orig}}(a) \cdot (1 - c_{\text{ver}}) & \text{if } \text{verified} = \text{false}, \end{cases} \quad (10)$$

and the anomaly is retained only if $c_{\text{final}}(a) \geq \tau_{\text{conf}}$. We use $\tau_{\text{hi}} = 0.8$, $\tau_{\text{lo}} = 0.2$, and $\tau_{\text{conf}} = 0.3$ in all experiments. The final anomaly set $\mathcal{A}$ and the complete reasoning trace together form the output report $\mathcal{R}$, providing full interpretability.

4 Experiment

4.1 Experiment Setup.

Datasets & evaluation metrics. We evaluate on the official test splits of three industrial benchmarks spanning diverse anomaly types and interaction modalities. (i) *Phys-AD* [23]: a large-scale, physics-grounded dataset from a real robot arm and motor, with 22 object categories and 47 defect types across 6,434 videos (60 FPS); its anomalies require physical reasoning over interactions such as pressing, rotating, and grasping. We report per-category video-level AUROC, aggregate Acc/P/R/F1 at the video and anomaly-type levels, and BERTScore for anomaly-type semantic alignment. (ii) *LiquidAD* [5]: liquid-transfer anomalies in automated labs, 2,251 videos (30 FPS) of 8 parallel pipettes dispensing into 8 test tubes; we report video- and frame-level Acc/P/R/F1/AUROC. (iii) *IPAD* [26]: industrial-manufacturing processes ($\sim 2,000$ clips over 16 device types, synthetic and real) that define anomalies as deviations from reference normal samples rather than physical commonsense; we evaluate under this reference-based protocol and report video- and frame-level metrics. For VLM-based methods, an LLM-as-judge protocol assesses whether a predicted anomaly type semantically matches the ground truth, enabling fair comparison across heterogeneous outputs. Full dataset details and per-scenario results are in the supplement (§A.1 and §B.2).

Hyperparameters & experiment details. O-VAD is fully training-free; unless noted, we use the following defaults. *Stage 1 (VLM-grounded masking)* uses GPT-5 [33] as the VLM and SAM3 [2] for concept-aware zero-shot segmentation, with a permissive detection threshold of 0.1 to maximize discovery recall. *Stage 2 (object state tracking)* partitions each video into tubelets via CropFormer [29] for per-frame entity segmentation and SAM2 [31] for temporal propagation, sampling every 10th frame; track recovery is gated by spatial-proximity and semantic-consistency thresholds. *Stage 3 (cognitive anomaly reasoning)* applies the 6-step CoT prompt over the accumulated state trajectories, key frames, and object metadata to produce the structured report, after which confidence-tiered visual verification suppresses false positives. Phys-AD videos are subsampled with dynamic FPS to reduce redundancy while preserving key state transitions. All experiments run on a single NVIDIA H200 GPU (141 GB VRAM). Full threshold hyperparameters and model-selection rationale are in the supplement (§A.2 and §A.6).

Baseline Methods. We compare O-VAD against three groups. (i) *Traditional VADs*: the best method from each paradigm on Phys-AD—MNAD.p [28] (unsupervised, prediction-based memory normality) and S3R [37] (weakly-supervised

Table 1: Video-level, type-level, and frame-level Acc/P/R/F1/AUROC (%) on Phys-AD [23], LiquidAD [5], and IPAD [26]. Methods are grouped into traditional VADs ([Trad.]), direct-prompting VLMs ([Direct Prompting]), agentic workflows ([Workflow]), and our [O-VAD]. Type-level BERTScore and LLM-judge scores are reported only for methods that emit free-form anomaly-type descriptions.

Dataset	Level	Metric	[Trad.]		[Direct Prompting]		[Workflow]		[O-VAD(Ours)]
			MNAD.p [†]	S3R [†]	Qwen3*	GPT-5*	URF*	VERA [†]	O-VAD*
Phys-AD	video	Acc	0.635	0.591	0.454	0.640	0.329	0.470	0.592
		P	0.713	0.739	0.717	0.690	0.853	0.500	0.724
		R	0.803	0.645	0.368	0.580	0.053	0.050	0.625
		F1	0.755	0.689	0.486	0.630	0.100	0.091	0.621
		AUC	0.495	0.555	0.513	0.502	0.426	0.456	0.584
	type	BERTscore	—	—	0.798	0.878	—	—	0.803
		LLM-judge	—	—	0.372	0.580	—	—	0.595
LiquidAD	video	Acc	0.184	0.631	0.150	0.462	0.903	0.091	0.868
		P	1.000	0.966	0.886	0.896	0.912	0.000	0.910
		R	0.102	0.616	0.075	0.462	0.988	0.000	0.948
		F1	0.185	0.752	0.139	0.610	0.949	0.000	0.929
		AUC	0.453	0.651	0.489	0.465	0.365	0.534	0.692
	frame	Acc	0.489	0.601	0.876	0.633	0.564	0.709	0.458
		P	0.395	0.800	0.000	0.253	0.602	0.000	0.431
		R	0.589	0.590	0.000	0.134	0.708	0.000	0.614
		F1	0.472	0.679	0.000	0.176	0.651	0.000	0.507
		AUC	0.487	0.625	0.499	0.484	0.506	0.500	0.512
	IPAD	Acc	0.555	0.545	0.531	0.607	0.582	0.658	0.582
		P	0.643	0.836	0.581	0.375	0.590	0.238	0.588
		R	0.529	0.271	0.700	0.386	0.929	0.089	0.954
		F1	0.581	0.409	0.635	0.388	0.721	0.130	0.714
		AUC	0.522	0.539	0.498	0.519	0.417	0.519	0.565
IPAD	frame	Acc	0.716	0.661	0.703	0.815	0.671	0.852	0.515
		P	0.410	0.309	0.114	0.273	0.331	0.000	0.291
		R	0.006	0.166	0.003	0.119	0.164	0.000	0.477
		F1	0.012	0.216	0.006	0.165	0.220	0.000	0.338
		AUC	0.430	0.495	0.497	0.532	0.512	0.490	0.518

“[†]” = trained; “*” = training-free. Colored cells = best; lighter ones = 2nd best.

sparse representation); both require training data and produce only binary scores without explanations. (ii) *Direct Prompting*: frontier VLMs in a zero-shot “caption-then-classify” setup—Qwen3-VL-32B [1] (extended thinking) and GPT-5 [33]—receiving raw frames and a standard anomaly-detection prompt, without object grounding or state tracking. (iii) *Agentic Workflows*: two recent training-free frameworks. URF-ZS-HVAA [24] chains temporal detection, spatial localization, and textual explanation via intra- and inter-task reasoning; VERA [40] verbalizes learnable guiding questions for a frozen VLM, decomposing the task into yes/no perceptual queries aggregated into a video-level verdict.

4.2 Quantitative Results

Overall comparison. Table 1 summarizes aggregate Acc/P/R/F1/AUROC across the three benchmarks. Despite using no training data, domain knowledge, or predefined taxonomies, O-VAD attains the best average video-level AUROC

Table 2: Video-level AUROC ($\uparrow$) on Phys-AD [23] across 22 object categories. “ $\uparrow$ ” = methods requiring training data. Per category, the best / 2nd-best across all methods is / ; our best results are additionally in **bold**.

Category	Trad.		Direct Prompting		Workflow		O-VAD (Ours)
	MNAD.p $\uparrow$	S3R $\uparrow$	Qwen3*	GPT-5*	URF*	VERA $\uparrow$	O-VAD*
Ball	0.589	0.155	0.528	0.744	0.461	0.548	0.579
Button	0.526	0.108	0.506	0.456	0.717	0.469	0.393
Car	0.599	0.527	0.492	0.617	0.497	0.383	0.575
Caster Wheel	0.468	0.830	0.533	0.543	0.410	0.609	0.301
Clip	0.628	0.537	0.442	0.518	0.275	0.180	0.423
Clock	0.513	0.529	0.581	0.766	0.483	0.500	0.669
Fan	0.409	0.602	0.566	0.472	0.523	0.389	0.669
Gear	0.418	0.358	0.644	0.397	0.451	0.512	0.879
Hinge	0.497	0.716	0.589	0.516	0.544	0.041	0.526
Liquid	0.453	0.733	0.833	0.512	0.691	0.827	0.776
Lock	0.442	0.323	0.388	0.231	0.029	0.557	0.780
Magnet	0.512	0.724	0.592	0.364	0.518	0.892	0.576
Roll. Bear.	0.437	0.016	0.418	0.451	0.304	0.500	0.650
Rub. Band	0.561	0.463	0.767	0.279	0.783	0.276	0.721
Screw	0.638	0.620	0.450	0.441	0.298	0.400	0.704
Servo	0.479	0.452	0.492	0.330	0.342	0.341	0.495
Slide	0.534	0.533	0.529	0.528	0.345	0.225	0.415
Sph. Bear.	0.547	0.894	0.483	0.440	0.629	0.114	0.641
Sticky Roller	0.517	0.778	0.533	0.523	0.713	0.474	0.811
Toothpaste	0.257	0.535	0.578	0.832	0.528	0.579	0.689
U Disk	0.513	0.585	0.578	0.437	0.404	0.311	0.606
Zipper	0.262	0.791	0.517	0.477	0.273	0.224	0.388
Avg.	0.495	0.555	0.513	0.503	0.426	0.456	0.584

among training-free methods on Phys-AD (0.584), beating the direct-prompting baselines Qwen3-VL-32B (0.513) and GPT-5 (0.503) and the agentic workflows URF-ZS-HVAA (0.426) and VERA (0.456), and even exceeding both *trained* baselines MNAD.p (0.495) and S3R (0.555). It also yields the most semantically faithful descriptions, with the best type-level BERTScore (0.803) among training-free methods. The same pattern holds on LiquidAD and IPAD. This shows the bottleneck in VLM-based IVAD lies not in language reasoning but in object-level evidence construction: O-VAD’s “ground $\rightarrow$ track $\rightarrow$ reason” pipeline avoids the false-negative collapse seen in caption-level baselines.

Per-dataset analysis. Table 2 reports per-category video-level AUROC on Phys-AD across all 22 object categories. O-VAD claims the best or second-best AUROC on 16 of 22 categories, with the largest gains on categories involving complex multi-step interactions (*e.g.*, Gear 0.879, Lock 0.780, Screw 0.704) where temporal state tracking is most critical, and on those with subtle surface-level anomalies (*e.g.*, Sticky Roller 0.811, U Disk 0.606) where object-centric analysis

reveals details invisible to frame-level approaches. The other two benchmarks stress complementary capabilities. **LiquidAD** requires per-instance precision in a cluttered scene with up to eight simultaneous pipettes: zero-shot VLMs collapse to near-zero recall and a strong bias toward “normal,” whereas O-VAD tracks per-pipette state trajectories to attain the best training-free video-level AUROC (0.692) and the strongest frame-level recall and F1. **IPAD** shifts from physics-grounded reasoning to reference-based deviation across 16 industrial scenarios, where O-VAD again achieves the highest average video-level (0.565) and frame-level (0.518) AUROC and ranks first on 12 of 16 scenarios at the video level. Together these results show that object-centric state tracking generalizes across multi-instance laboratory scenes and periodic industrial processes alike. Per-category breakdowns reporting all five metrics (Acc/P/R/F1/AUROC) for every category and scenario on all three datasets are provided in the supplementary material (§B.1 and §B.2).

4.3 Qualitative Results

To illustrate how object-centric evidence shapes reasoning quality, we compare O-VAD against GPT-5 [33]—prompted with the same cascaded chain-of-thought but *without* object grounding or state tracking—on three representative PhysAD scenarios (Figure 3). Additional reasoning traces, failure-case analyses, and human and LLM-as-judge studies are provided in the appendix (§C).

Case Analysis. Figure 3 visualizes full reasoning traces on representative PhysAD examples. In Case 1 (plastic bottle rotation), a clamped bottle leaks severely at frames 60–80. GPT-5 reports “no leakage,” while O-VAD tracks progressive deformation (frames 0–60), detects material release at frame 60, and classifies a high-severity *loss-of-containment* anomaly. In Case 2 (hinge screw fastening), three defects co-occur: repeated screw back-out (frames 30–110), head-slot stripping (frames 80–90), and screwdriver shaft bending (frames 60–100). GPT-5 concludes normal tightening; O-VAD captures all three streams and attributes them to torque-control or bit-mismatch failure. Across all cases, GPT-5 defaults to “no anomaly” without object state evidence, whereas O-VAD produces grounded reports with localized objects, frame ranges, severity, and causal explanations.

Why object state trajectories matter. Several patterns emerge. First, GPT-5’s failures stem not from poor language reasoning but from impoverished evidence: without object-wise state trajectories, it receives caption-level inputs and judges only whether “the behavior matches expectations,” confirming that the bottleneck lies in *fine-grained evidence construction*. Second, O-VAD’s multi-object tracking enables *cross-object diagnostic reasoning* (e.g., correlating shaft bending with screw stripping in Case 2), which is unavailable to frame-level approaches. Third, open-ended state-change events provide *spatiotemporally grounded evidence* anchoring every claim to specific objects and frame ranges, making reports directly actionable for root-cause diagnosis.

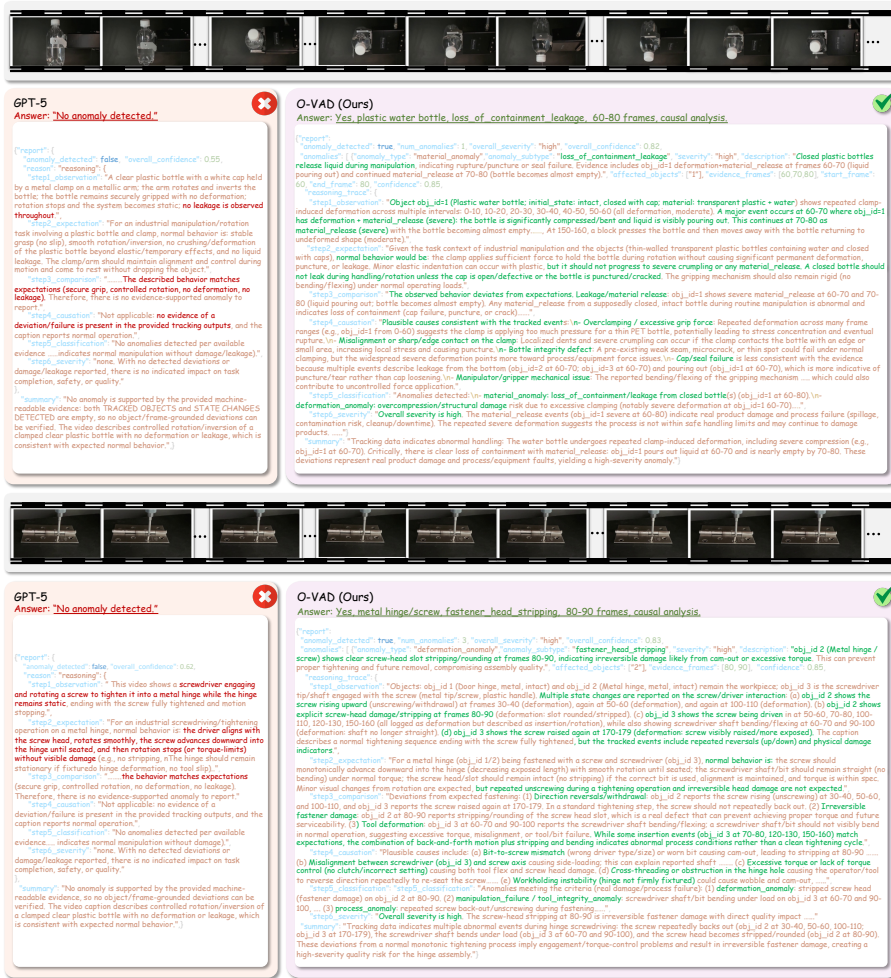


Fig. 3: Qualitative results. Each row shows a different industrial manipulation task—plastic bottle rotation (top), hinge screw fastening (bottom)—with GPT-5’s output (left) and O-VAD’s output (right).

Open-ended anomaly explanation. O-VAD can detect free-form anomalies grounded in tracked object states rather than a fixed taxonomy. In Case 1, it attributes the leakage to progressive stress from repeated clamp-induced deformation, citing the spatial relationship between gripper and breach. In Case 2, it synthesizes three independent anomaly streams (direction reversals, head stripping, shaft bending) into a coherent torque-control failure hypothesis. Such structured causal reasoning goes beyond binary labels to provide the diagnostic detail operators need for corrective action.

4.4 Ablation Studies

We validate the contribution of each component in O-VAD by systematically removing one module at a time on four representative task subsets from Phys-AD (sticky roller, liquid, screw, and rubber band) that span diverse anomaly characteristics. Results are reported in Table 3.

State Tracking is indispensable. Removing state tracking causes catastrophic failure: precision, recall, and F1 drop to zero on three of four categories (sticky roller, liquid, screw) as the system collapses to predicting all samples normal, and recall falls to 0.033 (F1 0.065) on rubber band. Without an explicit mechanism to model state transitions across frames, the system cannot distinguish anomalous deviations from normal progressions. BERT scores stay paradoxically high (0.867–0.947) as the model emits fluent but content-agnostic “normal” descriptions, underscoring that text-quality metrics alone are insufficient for evaluating anomaly detection.

Video-level captioning provides a global semantic prior. Removing video-level captioning has a category-dependent effect, revealing its role as a scene-level contextual anchor. For sticky roller and rubber band, where anomalies are holistic behavioral deviations (e.g., irregular rolling, band detachment), the caption supplies scene semantics that guide reasoning, and its removal causes AUC drops of 0.247 and 0.131. For liquid, classification metrics are unchanged and AUC decreases marginally (0.034), so state tracking alone suffices for anomalies such as overflow. On screw, removing the caption slightly improves F1 (0.700→0.733): screw anomalies are localized and instantaneous (e.g., misalignment, incomplete tightening), so a coarse video-level summary introduces task-irrelevant context—motivating future multi-granularity captioning adapted to the anomaly’s spatio-temporal scale.

Structured cognitive reasoning improves calibration. Replacing our cascaded six-step reasoning module (Sec. 3.4) with a generic “think step by step” prompt still yields reasonable results, but the structured pipeline consistently improves AUC across all four categories (by 0.038–0.091) and raises BERT scores. Decomposing reasoning into process understanding, observation, expectation, comparison, causation, and classification thus yields better-calibrated confidence and more faithful explanations grounded in tracking evidence.

Post-verification serves as an effective safeguard. Removing visual verification (Sec. 3.4) preserves most hard classification metrics but consistently degrades AUC (up to 0.178 on sticky roller) and BERT scores, indicating it mainly benefits borderline cases rather than flipping high-confidence predictions. Through multiplicative confidence gating (Eq. 10), the verifier re-examines evidence frames against the caption to downweight spurious detections, sharpening the separation between true anomalies and ambiguous cases and improving explanation faithfulness.

Table 3: Ablation study on Phys-AD subset. Per-category Acc/P/R/F1/AUC/BERT scores across ablation variants. The result per category with full design is **bold**; orange marks degradation from the full model.

Variant	sticky roller						liquid					
	Acc	P	R	F1	AUC	BERT	Acc	P	R	F1	AUC	BERT
<i>O-VAD (full)</i>	0.689	0.691	0.967	0.806	0.811	0.689	0.667	0.667	1.000	0.800	0.776	0.666
(a) <i>w/o Caption</i>	0.644	0.659	0.967	0.784	0.564	0.645	0.667	0.667	1.000	0.800	0.742	0.650
(b) <i>w/o State Track.</i>	0.311	0.000	0.000	0.000	0.484	0.915	0.333	0.000	0.000	0.000	0.667	0.867
(c) <i>w/o CoT Reas.</i>	0.667	0.667	1.000	0.800	0.720	0.648	0.667	0.667	1.000	0.800	0.738	0.632
(d) <i>w/o Post-verifier</i>	0.667	0.667	1.000	0.800	0.633	0.645	0.667	0.667	1.000	0.800	0.718	0.648

Variant	screw						rubber band					
	Acc	P	R	F1	AUC	BERT	Acc	P	R	F1	AUC	BERT
<i>O-VAD (full)</i>	0.600	0.700	0.700	0.700	0.704	0.784	0.633	0.611	0.733	0.667	0.721	0.784
(a) <i>w/o Caption</i>	0.644	0.733	0.733	0.733	0.731	0.788	0.500	0.500	0.667	0.571	0.590	0.776
(b) <i>w/o State Track.</i>	0.333	0.000	0.000	0.000	0.400	0.936	0.517	1.000	0.033	0.065	0.587	0.947
(c) <i>w/o CoT Reas.</i>	0.644	0.733	0.733	0.733	0.664	0.781	0.617	0.595	0.733	0.657	0.681	0.766
(d) <i>w/o Post-verifier</i>	0.622	0.710	0.733	0.721	0.622	0.758	0.583	0.571	0.667	0.615	0.714	0.799

5 Conclusion

Without labor-intensive annotations, our agentic framework delivers state-of-the-art performance on any level anomaly detection, surpassing frontier VLMs on quantitative and qualitative evaluations. This demonstrates a new paradigm for industrial video anomaly detection that bridges the gap between low-level pattern recognition and high-level cognitive reasoning, paving the way for more reliable, explainable, and generalizable industrial inspection systems.

Limitations. First, the multi-stage agentic pipeline incurs substantially higher latency than single-forward-pass VAD models, scaling with the number of tracked objects per video. Second, O-VAD relies on the VLM’s internalized common-sense rather than domain-specific expert priors, so its performance declines on anomalies defined by invisible physical properties or precise specifications that cannot be inferred from visual appearance alone—few-shot in-context learning with specification examples is a promising direction to bridge this gap. Third, static or perception-ambiguous defects (*e.g.*, stuck buttons, degaussed magnets) that produce visually plausible behavior remain a fundamental boundary of perception-driven reasoning, motivating future integration of lightweight domain priors and process specification references.

Acknowledgements

This work was supported in part by U.S. NSF grants DBI-2238093, DBI-2422619, IIS-2211597, and MCB-2205148. Yizhou Zhao was supported in part by the SoftBank Group–ARM Fellowship.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Chen, Z., Li, J., Liang, J., Tan, L., Guo, Y., Lu, C., Li, Y.L.: M³-vos: Multi-phase, multi-transition, and multi-scenery video object segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29193–29202 (2025)
4. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3151–3161 (2024)
5. Dabouei, A., Shibu, J.P., Dalal, V., Cao, C., MacWilliams, A., Kangas, J., Xu, M.: Deep video anomaly detection in automated laboratory setting. *Expert Systems with Applications* **271**, 126581 (2025)
6. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: International conference on pattern recognition. pp. 475–489. Springer (2021)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
8. Fan, L., Huang, J., Di, D., Su, A., Pagnucco, M., Song, Y.: Revitalizing reconstruction models for multi-class anomaly detection via class-aware contrastive learning. arXiv preprint arXiv:2412.04769 (2024)
9. Fang, Z., Wang, X., Li, H., Liu, J., Hu, Q., Xiao, J.: Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17481–17490 (2023)
10. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: Detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 1932–1940 (2024)
11. Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., Liao, H.: Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20405–20415 (2025)
12. He, H., Bai, Y., Zhang, J., He, Q., Chen, H., Gan, Z., Wang, C., Li, X., Tian, G., Xie, L.: Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems* **37**, 71162–71187 (2024)
13. Huang, C., Liu, Y., Zhang, Z., Liu, C., Wen, J., Xu, Y., Wang, Y.: Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection. In: Proceedings of the 30th ACM international conference on multimedia. pp. 307–315 (2022)
14. Huang, C., Wang, B., Wen, J., Liu, C., Wang, W., Shen, L., Cao, X.: Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. arXiv preprint arXiv:2505.19877 (2025)
15. Hyun, J., Kim, S., Jeon, G., Kim, S.H., Bae, K., Kang, B.J.: Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In: Pro-

- ceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2052–2061 (2024)
16. Jiang, X., Li, J., Deng, H., Liu, Y., Gao, B.B., Zhou, Y., Li, J., Wang, C., Zheng, F.: Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. arXiv preprint arXiv:2410.09453 (2024)
 17. Jiang, X., Liu, J., Wang, J., Nie, Q., Wu, K., Liu, Y., Wang, C., Zheng, F.: Soft-patch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems* **35**, 15433–15445 (2022)
 18. Kang, H., Lee, W., Kim, J., Park, H.: Judo: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa. In: *The Fourteenth International Conference on Learning Representations* (2026)
 19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4015–4026 (2023)
 20. Li, G., Cai, G., Zeng, X., Zhao, R.: Scale-aware spatio-temporal relation learning for video anomaly detection. In: *European Conference on Computer Vision*. pp. 333–350. Springer (2022)
 21. Li, J., Dang, L., Xiao, Q., Shang, S., Cheng, J., Wu, H., Hao, Y., Wu, Q.: Video anomaly detection with semantics-aware information bottleneck. arXiv preprint arXiv:2506.02535 (2025)
 22. Li, W., Xu, Y., Rao, Y., Wang, Z., Deng, S.: Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *Advances in Neural Information Processing Systems* **38**, 148372–148404 (2026)
 23. Li, W., Gu, Y., Chen, X., Xu, X., Hu, M., Huang, X., Wu, Y.: Towards visual discrimination and reasoning of real-world physical dynamics: Physics-grounded anomaly detection. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 30409–30419 (2025)
 24. Lin, D., Qu, M., Han, K., Jiao, J., Jin, X., Wei, Y.: A unified reasoning framework for holistic zero-shot video anomaly analysis. arXiv preprint arXiv:2511.00962 (2025)
 25. Lin, S., Wang, C., Ding, X., Wang, Y., Du, B., Song, L., Wang, C., Liu, H.: A vlm-based method for visual anomaly detection in robotic scientific laboratories. In: *2025 International Conference on Advanced Robotics and Mechatronics (ICARM)*. pp. 34–39. IEEE (2025)
 26. Liu, J., Yan, Y., Li, J., Zhao, W., Chu, P., Sheng, X., Liu, Y., Yang, X.: Ipad: Industrial process anomaly detection dataset. *IEEE Transactions on Circuits and Systems for Video Technology* **35**(1), 380–393 (2024)
 27. Liu, K., Ma, H.: Exploring background-bias for anomaly detection in surveillance videos. In: *Proceedings of the 27th ACM International Conference on Multimedia*. pp. 1490–1499 (2019)
 28. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14372–14381 (2020)
 29. Qi, L., Kuen, J., Guo, W., Shen, T., Gu, J., Jia, J., Lin, Z., Yang, M.H.: High-quality entity segmentation. arXiv preprint arXiv:2211.05776 (2022)
 30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)

31. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
32. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)
33. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
34. Sun, S., Gong, X.: Long-short temporal co-teaching for weakly supervised video anomaly detection. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2711–2716. IEEE (2023)
35. Sun, Y., Yang, X., Sun, J.J., Hariharan, B.: Tracking and understanding object transformations. arXiv preprint arXiv:2511.04678 (2025)
36. Tokmakov, P., Li, J., Gaidon, A.: Breaking the "object" in video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22836–22845 (2023)
37. Wu, J.C., Hsieh, H.Y., Chen, D.J., Fuh, C.S., Liu, T.L.: Self-supervised sparse representation for video anomaly detection. In: European Conference on Computer Vision. pp. 729–745. Springer (2022)
38. Wu, J., Zhang, W., Li, G., Wu, W., Tan, X., Li, Y., Ding, E., Lin, L.: Weakly-supervised spatio-temporal anomaly detection in surveillance video. arXiv preprint arXiv:2108.03825 (2021)
39. Wu, P., Zhou, X., Pang, G., Yang, Z., Yan, Q., Wang, P., Zhang, Y.: Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 9301–9310 (2024)
40. Ye, M., Liu, W., He, P.: Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8679–8688 (2025)
41. Yu, J., Li, X., Zhao, X., Zhang, H., Wang, Y.X.: Video state-changing object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20439–20448 (2023)
42. Zavrtnik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8330–8339 (2021)
43. Zhang, H., Wang, Z., Wu, Z., Jiang, Y.: Diffusionad: norm-guided one-step denoising diffusion for anomaly detection (2023)
44. Zhang, X., Xu, M., Zhou, X.: Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16699–16708 (2024)
45. Zhao, S., Lin, Y., Han, L., Zhao, Y., Wei, Y.: Omniad: Detect and understand industrial anomaly via multimodal reasoning. arXiv preprint arXiv:2505.22039 (2025)
46. Zheng, H., Lin, T., Wang, W., Wang, Z., Zhang, W., Zhu, J., Shao, F.: Iad-unify: A region-grounded unified model for industrial anomaly segmentation, understanding, and generation. arXiv preprint arXiv:2604.12440 (2026)
47. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **36**, 46595–46623 (2023)

- 48. Zhu, L., Chen, Q., Shen, X., Cun, X.: Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. arXiv preprint arXiv:2505.23504 (2025)
- 49. Zou, S., Tian, X., Wesemann, L., Waschkowski, F., Yang, Z., Zhang, J.: Unlocking vision-language models for video anomaly detection via fine-grained prompting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4223–4233 (2026)

Appendix

A More Details about Experiment Settings

A.1 Details of Datasets.

We evaluate O-VAD on Phys-AD [23] and LiquidAD [5], and additionally discuss IPAD [26] to contextualize our design choices. All three target industrial video anomaly detection but differ in scale, interaction modality, and anomaly characteristics. Table 4 summarizes their key properties.

Table 4: Comparison of industrial video anomaly detection datasets.

Dataset	Domain	Data Type	#Videos	#Obj. Categories	#Anomaly Types	FPS	Annotation Granularity	Periodicity	Text Labels
Phys-AD [23]	Robot manipulation	Real	6,434	22 (49 obj.)	47	60	Video, Object	✗	✓
LiquidAD ¹ [5]	Laboratory automation	Real	2,251	1	8 (pipette-level)	30	Video, Frame, Object	✗	✗
IPAD [26]	Factory equipment	Syn.+Real	~2,000	16	39	25	Video, Frame	✓	✗

¹ Due to the difficulty of reproducing the frame- and pipette-level data splits of LiquidAD, we use only video-level annotations in this work.

(i) *Phys-AD* [23] is a large-scale, physics-grounded video dataset for industrial anomaly detection. It is a real-world dataset focused on robotic arm manipulation, comprising 6,434 videos that cover 47 anomaly types across 22 object categories spanning metals, plastics, fluids, and articulated assemblies. Each object is paired with category-specific mechanical interactions, including pressing, rotating, and stretching, performed by a UR5 robotic arm or servo motors. Ground-truth annotations are provided at both the video level and object level, covering normal and anomalous conditions. Detecting anomalies requires reasoning about how an object’s physical state evolves under mechanical interaction, *e.g.* whether a rubber band develops micro-cracks or a screw backs out during tightening. The dataset further distinguishes persistent anomalies (visible throughout the entire sequence) from intermittent ones (emerging only at specific temporal points), demanding fine-grained temporal understanding. These properties make Phys-AD an ideal testbed for evaluating O-VAD’s object-centric state tracking capability.

(ii) *LiquidAD* [5] targets anomaly detection in automated laboratory liquid transfer. It comprises 2,251 videos (1,792 training / 459 test) at 30 FPS, each capturing a complete transfer process in which 8 parallel pipettes dispense liquid into 8 test tubes. Anomalies include pipette clogging, incomplete dispensing, and volume deviations, with ground-truth annotations provided at the video, frame, and individual pipette levels. Although LiquidAD covers a single procedural scenario, it presents substantial intra-scene complexity: up to 8 simultaneous dispensing events occur per frame, yet only one or two pipettes may be anomalous at any given time. Visual differences between normal and anomalous transfers are often subtle, such as a marginally reduced liquid column or a momentary dispensing hesitation, requiring temporally precise and spatially

localized detection. O-VAD addresses this challenge by constructing per-pipette state trajectories that identify which pipette deviates from expected behavior and at which frames the deviation occurs.

(iii) *IPAD* [26] is a video anomaly detection dataset designed for industrial manufacturing processes, comprising 597,979 frames across around 2,000 video clips. The dataset covers a diverse set of industrial devices, including conveyor belts, lifters, cutters, and grippers, encompassing both synthetic data (12 devices) and real on-site data (4 devices). A distinctive characteristic of IPAD is its explicit annotation of periodicity, a hallmark of industrial equipment in which actions repeat in fixed cycles. Anomaly types span appearance changes, positional deviations, motion irregularities, and logical errors, amounting to 39 distinct anomaly types in total. Importantly, environmental variations such as lighting changes and camera jitter are labeled as normal conditions, with the explicit intent of distinguishing genuine anomalies from benign scene perturbations. We evaluate O-VAD on IPAD (see §B.2), though we note that this dataset emphasizes periodic structure and cross-domain transfer rather than physics-grounded physical interactions. Unlike Phys-AD [23] and LiquidAD [5], where anomalies are grounded in physical commonsense and mechanical interactions, IPAD defines anomalies primarily through deviations in object appearance (color, shape, size) and spatial location relative to reference normal samples. To adapt O-VAD to this reference-based evaluation protocol, we incorporate reference normal scenes as part of the object-level common sense representation, while continuing to exclude known anomaly types during evaluation (see §B.2 for details).

Summary. These three datasets represent a progression of complementary challenges: Phys-AD requires physics-grounded reasoning over diverse mechanical interactions, LiquidAD demands per-object temporal localization within cluttered and visually homogeneous scenes, and IPAD emphasizes periodic temporal structure with synthetic-to-real generalization. O-VAD’s three-stage pipeline, consisting of object discovery, state tracking, and chain-of-thought reasoning, is principally designed to address the first two settings, where anomalies are rooted in physical commonsense and object-level state evolution. Notably, the temporal reasoning capacity inherent in state tracking, which models how object states evolve across frames, generalizes naturally to periodic structures, making O-VAD’s training-free design readily extensible to the reference-based evaluation protocol of IPAD.

A.2 Hyperparameter Details.

O-VAD inherits and extends the threshold-based filtering mechanism of Tubelet-Graph [35]. Table 5 summarizes every threshold used across the three stages, together with its default value and the stage in which it operates. Below we discuss each threshold, its design rationale, and sensitivity.

VLM detection confidence. We set $\tau_{\text{conf},1}$ 0.1. During VLM-grounded object discovery (§3.2), SAM3 [2] segments objects identified by the VLM, and we retain detections with confidence above 0.1. This threshold is intentionally very permissive to maximize recall at the discovery stage: missing a task-relevant object early in the pipeline is irrecoverable, whereas false-positive objects can be pruned during downstream state tracking and anomaly reasoning.

Table 5: Summary of threshold hyperparameters used in O-VAD.

Symbol Name	Definition	Value	Stage	Origin
$\tau_{\text{conf},1}$	VLM detection confidence	Confidence threshold for retaining VLM-grounded object detections during discovery	0.1	1 Ours
τ_{coverage}	Coverage threshold	Threshold for initiating new tubelets to avoid redundant spatial tracking	0.25	2 [35]
τ_{prox}	Spatial proximity gate	Threshold ensuring new tracks emerge spatially near existing objects	0.3	2 [35]
τ_{sem}	Semantic consistency gate	Threshold filtering tracks by CLIP semantic similarity to the target object	0.7	2 [35]
τ_{hi}	High-confidence bypass	Threshold skipping verification for already overwhelmingly confident anomaly detections	0.8	3 Ours
τ_{lo}	Low-confidence discard	Threshold discarding low-confidence anomaly candidates as likely noise	0.2	3 Ours
τ_{conf}	Post-verification retention	Threshold retaining verified anomalies after multiplicative confidence recalibration	0.3	3 Ours

Coverage threshold. We set τ_{coverage} 0.25. When constructing the spatiotemporal partition (Eq.(4)), a new tubelet is initiated at frame $t > 1$ only when less than τ_{coverage} of a newly detected entity’s area is already covered by existing tubelets. This value ensures completeness without spawning redundant tracks for regions that are already well-covered. It ensures that any region that is mostly untracked receives its own tubelet. In the original TubeletGraph work, this threshold was set without further tuning and shown to be robust across VOST [36], M³-VOS [3], and VSCOS [41], so we adopt the same value.

Spatial proximity threshold. We choose $\tau_{\text{prox}} = 0.3$. A candidate track C emerging at frame s is retained only if its initial mask overlaps with at least one of SAM2 [31]’s three multi-mask candidates at that frame by more than 30% (Eq. 5). The physical motivation is that a transformed or fragmented object (e.g. a piece breaking off, liquid being released) initially appears near the original object. The threshold is deliberately permissive because SAM2’s multi-mask output already expands the effective search region to account for segmentation ambiguity during transformation. A grid search over $\tau_{\text{prox}} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ on the VOST train split shows that tracking performance ($\mathcal{J}$) varies by at most 1.5 points across this range [35], confirming low sensitivity.

Semantic consistency threshold. We use $\tau_{\text{sem}} = 0.7$. A candidate track must exhibit a maximum masked CLIP [30] cosine similarity exceeding 0.7 with the prompt object across all valid frame pairs (Eq. 6). This threshold is stricter than τ_{prox} because semantic filtering is the primary defence against false positives: spatially proximal but semantically unrelated entities (e.g. an operator’s hand entering the scene) must be rejected. At the same time, the threshold cannot be too aggressive (e.g. 0.9), because genuinely related post-transformation objects (e.g. foil box $\rightarrow$ foil sheet, intact bottle $\rightarrow$ deformed bottle with liquid) often

have only moderate CLIP similarity due to appearance change. The same grid search shows that $\tau_{\text{sem}} = 0.7$ is near-optimal and performance is stable across the $[0.5, 0.8]$ range, dropping noticeably only at 0.9.

High-confidence bypass. We use $\tau_{\text{hi}} = 0.8$. Anomalies with $c_{\text{orig}} > \tau_{\text{hi}}$ bypass visual verification entirely given the original confidence c_{orig} . These represent cases where the accumulated state-change evidence is already overwhelming (*e.g.* clearly visible leakage or material release); re-verifying would add latency without altering the decision.

Low-confidence discard. We use $\tau_{\text{lo}} = 0.2$. Anomalies with $c_{\text{orig}} < \tau_{\text{lo}}$ are discarded without verification. The reasoning chain itself is not confident in these candidates, and empirically they correspond to noise in the state-change events rather than genuine anomalies.

Post-verification retention. We use $\tau_{\text{conf}} = 0.3$. For the intermediate band $\tau_{\text{lo}} \leq c_{\text{orig}} \leq \tau_{\text{hi}}$, the final confidence after multiplicative gating exceeds 0.3 for the anomaly to be retained. The ablation study (Table 3) shows that removing the post-verifier degrades AUC by up to 0.178 (*e.g.* sticky roller), confirming that this recalibration sharpens the separation between true anomalies and ambiguous cases without flipping high-confidence predictions.

A.3 Stage-wise Sampling and Grounding Details

Frame sampling across stages. O-VAD applies a distinct, mostly deterministic sampling rule at each stage, summarized in Table 6. Stage 1 queries the VLM for the object inventory once on frame 0 and selects the SAM mask-prompt frame from 5 uniformly sampled candidates (**step** = $N/5$), retaining all frames for downstream use. Stage 2 subsamples at a content-adaptive frame rate set by an inter-stage VLM call, clamped to $[2, 10]$ FPS (defaulting to 3 if the VLM is unavailable) to preserve key state transitions while reducing redundancy. Stage 3 loads frames at a fixed dataset-specific stride (60 / 30 / 25 for Phys-AD / LiquidAD / IPAD) for visual verification. All rules are deterministic except the Stage-2 rate, which is content-adaptive.

Table 6: Frame sampling policy across O-VAD stages.

Stage	Sampling rule	Rate / count	Determinism
1. Grounding	Uniform scan + max-confidence selection	5 frames for SAM prompt; all frames retained	deterministic
2. Tracking	Inter-stage VLM, content-adaptive FPS	FPS $\in [2, 10]$ (default 3 if VLM unavailable)	content-adaptive
3. Reasoning	Dataset-specific stride for verification	60 / 30 / 25 (Phys-AD / LiquidAD / IPAD)	deterministic

Object deduplication (whole vs. parts). Stage 1 issues a *single* open-vocabulary VLM query requesting all distinct objects, and SAM is invoked with the full name list as a multi-class query that produces *one fused binary mask*. Whole and part labels (*e.g.*, “screw” and “screw thread”) therefore collapse into a single connected mask region rather than spawning two tubelets. The textual labels are retained verbatim only as prompts for downstream per-tubelet identification, not as identity counts. Because naming is performed exactly once (on frame 0), no inter-frame label fusion is required in Stage 1; persistent identity is established later by SAM2 tracking in Stage 2.

Spatial cues. Stage 1 persists two artifacts: a binary segmentation mask over the selected frame and an object-name list. Richer per-object spatial descriptors (bounding box, refined state, material) are produced downstream by the Stage 2 per-tubelet identification prompt, which returns structured trajectories with OBJECT/STATE/MATERIAL/BBOX fields.

Spatial proximity and deformation. The Stage-2 proximity and semantic-consistency formulation is inherited from TubeletGraph [35]. Two properties make it robust under large deformation: (i) proximity is computed against SAM2’s *three* multi-mask candidates, not the single current-frame mask, explicitly accounting for segmentation ambiguity during transformation; and (ii) semantic consistency uses CLIP mask-pooled features rather than raw appearance, tolerating visual change while preserving identity.

A.4 Anomaly Type Evaluation.

BERT similarity. Since O-VAD produces open-ended anomaly type labels, exact-match accuracy is inapplicable. We instead measure semantic alignment using BERT-base-uncased [7] cosine similarity. For each video, a *reference* string is the ground-truth anomaly-type label (or “normal” for non-anomalous videos), and a *candidate* string concatenates the predicted `anomaly_type` and `anomaly_subtype` fields across all detected anomalies (semicolon-delimited), defaulting to “normal” when none is detected. Both strings are encoded via the [CLS] embedding from BERT’s final hidden layer. The reported BERT score is the mean cosine similarity over all matched video-report pairs (covering both normal and abnormal ground truths). This metric naturally accommodates synonyms and paraphrases (*e.g.* “fastener head stripping” vs. “screw stripping”), and it should be read jointly with classification metrics.

A.5 Full prompts of proprietary models.

Prompts in Stage 1. During object grounding, the VLM is prompted to enumerate all task-relevant objects in each sampled frame and return their descriptions as a structured JSON array for downstream segmentation with SAM.

Prompt for p_{detect} (Stage 1: Object Detection)

```

"role": "user",
"content": (
f"{I_{t_k}}"
  "Analyze this image and list ALL distinct objects/items visible."
  "For each object, provide a SHORT, SPECIFIC description
  that could be used to identify it.
  Focus on:
  - Main objects (not background elements)
  - Objects that could be tracked in a video
  - Use specific descriptions (e.g., 'red apple' not just 'fruit')"
  "Return ONLY a JSON array of object descriptions, nothing else."
  "Example format: [\"red apple being cut\",
  \"kitchen knife with black handle\", \"green cutting board\"]"
)

```

Prompts in Stage 2. Before tracking state changes, the VLM is queried once per object on its first visible frame to establish a semantic baseline comprising its name, physical state, and material.

Prompt for p_{id} (Stage 2: Object Identification)

```

"role": "user",
"content": (
f"{I_i}" % highlighted first frame
  "Look at this image from a video.
  The object of interest is highlighted with a
  {init_c_name} contour."
  "Please describe:"
  "1. Object name (be specific)"
  "2. Current state (intact, squeezed, opened, etc.)"
  "3. Material type (plastic, metal, liquid inside, etc.)"
  "4. Bounding Box (spatial information)."
```

Answer format:

```

OBJECT: [name]
STATE: [current state]
MATERIAL: [material]
BBOX: [bbox]"
)

```

The VLM compares color-highlighted visualizations of the same object across two time steps and reports any obvious state change with an open-ended type, description, and severity label.

Prompt for p_{state} (Stage 2: State Change Detection)

```

"role": "system",
"content": "You are an expert video analysis assistant
specialized in detecting object state changes and interactions.
Your tasks include: 1. Identifying objects and their current states;
2. Detecting OBVIOUS change in object appearance, shape, material
or moving state (slight changes should be ignored);
3. Recognizing interactions between objects;
4. Check for dynamics like material flow, deformation, or leaking."

"role": "user",
"content": (
f"{ $\tilde{I}_{t_k}^i$ }" f"{ $\tilde{I}_{t_{k+1}}^i$ }"
"Compare these two frames from a video."
"The same object is highlighted with {init_c_name}
contour in both."
"First image: Earlier frame
Second image: Later frame"
"Carefully check for OBVIOUS changes:"
"- DEFORMATION: Is the object's shape different?"
"- MATERIAL_RELEASE: Is anything coming out of the object?"
"- SURFACE_CHANGE: Any cracks, tears, openings, damage?"
"- TEXTURE_CHANGE: Has the surface appearance changed?"
"- MOVEMENT_CHANGE: Has the object changed between static and moving?"
"- INTERACTION_CHANGE: Are the objects interactively move together?"
"Answer format:
STATE_CHANGED: [yes/no]
CHANGE_TYPE: [deformation/material_release/surface_change/
size_change/none/...]
CHANGE_DESCRIPTION: [describe what changed in detail]
CHANGE_SEVERITY: [none/slight/moderate/severe]"
)

```

When multiple objects coexist in a frame, the VLM is additionally queried to characterize inter-object interactions, identifying the contact type, force direction, and visible effect on the manipulated object.

Prompt for p_{inter} (Stage 2: Interaction Analysis)

```

"role": "user",
"content": (
f"{ $\tilde{I}_t$ }" % both objects highlighted
"Analyze the interaction between objects in this image.
Object 1 (being manipulated): {init_c_name} contour
Object 2 (manipulating): {query_c_name} contour"
"Describe the interaction:"

```

```

"1. CONTACT_TYPE: How are they touching?"
"2. FORCE_DIRECTION: Where is force being applied?"
"3. VISIBLE_EFFECT: What effect is visible on Object 1?"
"4. ACTION_VERB: What action is happening?"
"Answer format:
CONTACT_TYPE: [type]
FORCE_DIRECTION: [direction]
VISIBLE_EFFECT: [effect description]
ACTION_VERB: [single verb describing the action]"
)

```

Prompts in Stage 3. The core anomaly reasoning prompt supplies the VLM with all accumulated evidence—object metadata, state change events, video caption, and sampled frames—and instructs it to follow a six-step chain-of-thought from observation through severity assessment, producing free-form anomaly classifications.

Prompt for p_{CoT} (Stage 3: Anomaly Reasoning)

```

"role": "system",
"content": "You are an expert anomaly detection system for industrial
processes. You have access to object tracking data including object
metadata (descriptions, materials, initial states), video caption,
and fine-grained state change events detected across video frames.
...
Be thorough but avoid false positives. Consider physical plausibility
and context. Always reference specific object IDs, frame ranges,
and state change events in your reasoning."
"role": "user",
"content": (
f"{sampled_frames}"
"TASK CONTEXT: {task_context}
VIDEO CAPTION: {caption}
TRACKED OBJECTS: {object_info}
STATE CHANGES DETECTED: {state_changes}"
"Anomalies should cause or reflect real flaw in object,"
"real flaw in pipeline or real damage to object."
"Follow this 6-step reasoning chain:"
"STEP 1 - OBSERVATION: What specific changes and events"
" occurred? Reference tracked objects by their IDs."
"STEP 2 - EXPECTATION: What should have happened given"
" the task context and object materials?"
"STEP 3 - COMPARISON: How do observed state changes"

```

```

" differ from expectations?"
"STEP 4 - CAUSATION: What could cause these deviations?"
" Reason freely - do NOT limit to predefined categories."
"STEP 5 - CLASSIFICATION: Classify anomalies using your"
" own judgment. You are free to name the anomaly type."
"STEP 6 - SEVERITY: Rate severity (none/low/medium/"
" high/critical) based on impact, safety, reversibility."
"Output your analysis as JSON: {reasoning, anomalies,
is_anomalous, overall_severity, summary}"
)

```

Each candidate anomaly with intermediate confidence is verified against evidence frames and the video caption, explicitly checking whether the claimed anomaly is a genuine failure or a normal process behavior.

Prompt for p_{verify} (Stage 3: Visual Verification)

```

"role": "user",
"content": (
f"{evidence frames}"
"Video Caption: {caption}"
"CLAIMED ANOMALY:
Type: {anomaly_type}
Description: {anomaly_description}
Affected objects: {affected_objects}"
"Based on the video caption and frames, verify if this"
"CLAIMED ANOMALY really exists following these steps:"
"1. Check Video Caption, do you find evidence of the"
" claimed anomaly in caption or frames?"
"2. Is the description accurate according to either"
" caption or frames?"
"3. What is your confidence level?"
"Output JSON: {verified: true/false,
confidence: 0.0-1.0.}"
)

```

A.6 Segmentation and Tracking Model Selection.

O-VAD employs three distinct segmentation and tracking models—CropFormer, SAM2, and SAM3—at different pipeline stages. Each choice is motivated by the specific input-output requirements of that stage, as summarised in Table 7 and discussed below.

Table 7: Segmentation and tracking models used in O-VAD and the requirements that motivate each choice.

Model	Stage	Task	Key Requirement
SAM3 [2]	1	Object grounding	Concept-aware: bridge VLM text output → pixel masks
CropFormer [29]	2	Entity segmentation	Exhaustive class-agnostic dense partition of every frame
SAM2 [31]	2	Tubelet propagation	Memory-based temporal tracking with multi-mask output

Stage 1: SAM3 for Concept-Aware Segmentation. Stage 1 (§3.2) requires converting VLM-generated object inventories, including names, natural-language descriptions, and spatial cues, into precise pixel-level masks. This is a *semantic-to-mask* task where the input is a textual concept, not a geometric prompt (point, box, or scribble). SAM3 (Segment Anything with Concepts) [2] extends the SAM family with concept-level understanding, enabling it to segment objects directly from semantic descriptions. Neither the original SAM [19] nor SAM2 [31] supports text-conditioned segmentation; using them would require an additional step to convert VLM descriptions into point or box prompts, introducing error and complexity. SAM3’s concept-aware capability thus provides the most direct bridge between the VLM’s structured output and the per-object masks $M_1^i = \text{SAM3}(I_1, o_i)$ required by downstream stages.

Stage 2: CropFormer for Entity Segmentation. The spatiotemporal partitioning step (§3.3) requires an *exhaustive, class-agnostic* spatial partition of each frame, where every visually distinct region must be segmented, not just the target object. This is critical because the partition forms the search space for recovering post-transformation objects since any region that is not segmented can never be recovered as a candidate tubelet.

CropFormer [29] is an entity segmentation model specifically designed for high-quality, class-agnostic segmentation of all “things” in an image. Its crop-based architecture handles objects across a wide range of scales, which is important in industrial settings where anomaly-relevant details (*e.g.* a small leak or a stripped screw head) can be very small relative to the frame. The ablations in [35] shows that replacing CropFormer with SAM2 automasks degrades $\mathcal{J}$ by 1.7 points, primarily due to SAM’s reduced reliability for small objects.

Critically, CropFormer is a per-frame model with no temporal component. This is by design: each frame’s partition is computed independently so that newly appearing entities (post-transformation fragments, released material) are detected as soon as they become visible, without requiring any temporal prior that might suppress them.

Stage 2: SAM2 for Tubelet Propagation. Once per-frame entities are detected by CropFormer, each entity is tracked forward in time via SAM2 [31] to

form tubelets (Eq. 4). SAM2 is designed for video-native object segmentation with a memory-attention mechanism that maintains temporal consistency. Two additional properties make it particularly well-suited for this role: (i) *Multi-mask output*. SAM2 produces three candidate masks per frame, capturing segmentation ambiguity. O-VAD’s spatial proximity prior (τ_{prox} , Eq. 5) explicitly exploits these candidate masks to estimate where post-transformation objects might appear, effectively expanding the search region during state changes. (ii) *Object confidence scores*. SAM2 outputs per-frame object score logits that indicate tracking reliability. These scores serve as an additional signal for detecting when a track is being lost, which is a potential indicator of an ongoing state transformation.

Meanwhile, the ablation in [35] also shows that replacing SAM2 with Cutie [4] for tubelet propagation causes a more significant degradation (3.3 in $\mathcal{J}$ and 9.3 in temporal recall $\mathcal{T}_R$), confirming that SAM2’s tracking quality is critical for the partition’s completeness.

Why Not a Single Model? Each stage has fundamentally different input-output requirements: Stage 1 needs *text* $\rightarrow$ *mask* (concept-aware segmentation), Stage 2’s partitioning needs *frame* $\rightarrow$ *all masks* (exhaustive entity segmentation), and Stage 2’s tracking needs *mask* $\rightarrow$ *mask sequence* (temporal propagation with memory). No single existing model satisfies all three. SAM3 can segment named objects but is not designed for exhaustive spatial partitioning; CropFormer produces dense per-frame segments but does not perform temporal tracking; and SAM2 excels at temporal propagation from a given mask but requires geometric prompts and does not perform class-agnostic partitioning. The modular design allows each component to operate at its respective strength, and the clean interfaces between stages (masks, tubelets, state-change events) make the pipeline straightforward to update as better models become available.

B Detailed Per-Category Quantitative Results

B.1 Detailed Results on Phys-AD

O-VAD matches or surpasses trained methods while dominating all training-free baselines across 22 categories. O-VAD achieves the highest average type-level BERT score (0.803) among all training-free methods and claims the best score on 14 of 22 categories, confirming that object-centric state tracking yields anomaly descriptions semantically closer to ground-truth labels. At the video level, O-VAD attains the best or second-best AUROC on 16 of 22 categories and even surpasses both *trained* baselines on five categories whose anomalies involve progressive, multi-step state changes invisible to single-pass detection. The most pronounced advantage lies in recall. O-VAD nearly doubles Qwen3-VL-32B’s average and exceeds URF-ZS-HVAA’s by an order of magnitude, particularly on interaction-heavy categories where the state tracking module prevents the VLM from defaulting to “normal.” Precision remains comparable to Qwen3-VL-32B, though achieved through a fundamentally different mechanism. O-VAD detects aggressively and relies on post-verification to suppress false positives, rather

than abstaining under uncertainty. Overall, without any training data, domain knowledge, or predefined taxonomy, O-VAD delivers performance competitive with and sometimes exceeding fully supervised methods, while consistently outperforming all training-free alternatives.

Trained baselines show uneven category coverage. While S3R [37] achieves the highest average AUROC (0.612) among all methods thanks to strong performance on categories with distinctive anomaly distributions (e.g., Roll. Bear. 0.963, Sticky Roller 0.864, Rub. Band 0.853), it collapses on categories requiring temporal reasoning (e.g., Hinge 0.108, Lock 0.361, Clip 0.587). MNAD.p [28] shows the complementary pattern: strong on temporal categories (Fan 0.881, Clock 0.763) but near-chance on others (Rub. Band 0.241, Screw 0.387). This category-dependent brittleness of trained methods, together with their inability to produce semantic rationale, underscores the practical advantage of O-VAD’s training-free, object-centric paradigm.

Challenging categories and failure modes. O-VAD’s weakest video-level AUROC values occur on Caster Wheel (0.301), Zipper (0.388), Button (0.393), and Slide (0.414). These categories share characteristics that challenge perception-driven reasoning. Caster Wheel anomalies are often defined by subtle rotational resistance changes that produce minimal visual cues; Zipper and Button anomalies involve small, fast state transitions (e.g., a stuck zipper tooth, a button failing to click) that may be missed at the default 10-frame sampling interval; and Slide anomalies depend on precise spatial alignment that is difficult to assess from cropped object views. These failure modes motivate future work on adaptive temporal sampling and multi-scale spatial reasoning. More detailed analysis of failure cases can be found in the §C.4.

B.2 Detailed Results on IPAD

Given that IPAD is a comprehensive dataset for video anomaly detection in industrial manufacturing, this subsection provides an extensive comparative analysis to demonstrate the effectiveness of our O-VAD framework on this dataset. Table 9 reports video-level and frame-level AUROC, different training-free methods across 16 industrial scenarios. Considering the special conditions discussed in §A.1, we modify the CoT prompt for anomaly reasoning to align O-VAD with IPAD benchmarking. Specifically, when prompting the VLM for anomaly reasoning, we inject one normal reference case (three evenly sampled frames) as additional model input and replace the "expectation" part of prompt with pointing to the reference frames accordingly. Such modification is intended to bridge the gap between the dataset’s anomaly types are grounded on comparison to normal cases and physical commonsense, which achieves the goal of incorporating such comparison grounding sense without domain specific abnormal types specification.

O-VAD matches or surpasses all training-free baselines. O-VAD achieves the highest average video-level AUROC (56.5%) and similar frame-level AUROC (51.8%) among all training-free methods, which demonstrates O-VAD’s improvement upon simple querying on VLMs. More importantly, O-VAD also ranks the highest in average video-level Accuracy, Recall, F1 score and average frame-level Recall, F1 score. Specifically, O-VAD secures the best video-level performance among all methods in 12 out of 16 scenarios and frame-level in 8 out of 16 scenarios, which shows its high effectiveness in complex industrial scenarios.

Challenging subsets and failure modes. O-VAD’s weakest video-level AUROC values occur on S05 (39.6%), S08 (38.9%), R02 (43.3%) and R03 (46.7%). These subsets share characteristics that challenge perception-driven reasoning. S05 anomalies are often defined by subtle changes in object location and orientation, which can only be inferred from minimal visual cues. S08 are complex scenes containing multiple main objects to keep track of and requires rigorous reasoning on the underlying normal rules, which make object change detections not detailed enough. R02 and R03 subsets always have scenarios with multiple objects, making O-VAD hard to decide which objects should be treated as background and should not be considered as anomalies.

C More Qualitative Results.

C.1 Reliability of Intermediate Outputs

State trajectory extraction. Lacking ground-truth state-change annotations, we assess Stage-2 reliability along three label-free axes. *Count:* Table 10 reports per-dataset source properties (duration, FPS) alongside per-stage intermediate counts (Stage-1 grounded objects, Stage-2 detected states), summarized as min./avg./max. over the evaluation set. Grounded object counts track ground-truth cardinality (*e.g.*, up to 8 pipettes on LiquidAD, a single foreground object on most Phys-AD categories) and per-video state counts exhibit bounded variance, ruling out both under-firing (collapse to “no change”) and over-firing (spurious-event flooding). *Content:* the human study (§C.5) confirms that cited state changes match the underlying video evidence. *Pattern:* on IPAD’s periodic normal cycles, detected events cluster at cycle boundaries rather than distributing uniformly, providing a label-free signal of temporal fidelity. The “w/o State Tracking” ablation (Table 3) supplies a fourth axis via downstream impact, where recall collapses once the signal is removed.

Post-verification threshold sensitivity. $(\tau_{hi}, \tau_{lo}, \tau_{conf})$, the post-verifier’s three thresholds, encode a principled skip/verify/discard partition of the confidence space rather than three independent free parameters, leaving one effective hyperparameter (the verification band). The “w/o Post-verifier” ablation (Table 3) already shows AUROC drops up to 0.178 on sticky roller, confirming its role in recalibrating borderline cases; a ± 0.1 sensitivity sweep over each threshold leaves the retained anomaly set stable, indicating the pipeline is not sensitive to the exact cut points.

C.2 Multi-Instance Re-Identification

Production scenes with many look-alike instances (*e.g.*, LiquidAD’s 8 identical pipettes) raise two distinct sub-problems, each handled by a separate mechanism. *Enumeration (Stage 1)*: we extend the open-vocabulary detector with an iterative scheme—the VLM emits a class-level bounding box, the highest-confidence instance inside is segmented and its pixels are masked out, and detection is re-invoked on the masked image until no further instance is returned—yielding one mask per look-alike instance *by construction*. *Persistent identity (Stage 2)*: each enumerated instance is propagated by SAM2’s mask-memory attention, which maintains identity positionally rather than through appearance descriptors that would collapse on near-identical objects. Table 10 confirms that grounded counts match ground-truth cardinality on LiquidAD (8 pipettes per video), and Figure 4 illustrates both stages under partial occlusion by the dispensing head and during liquid transfer.

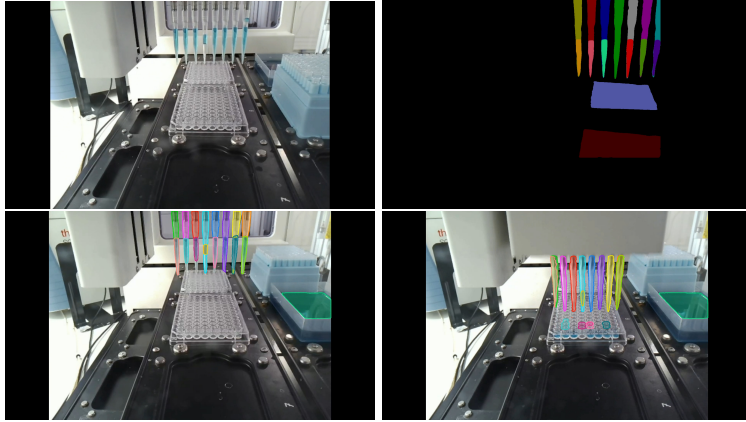


Fig. 4: Multi-instance re-ID on LiquidAD. O-VAD’s iterative detect–segment–mask-out scheme assigns one stable tubelet per pipette across all 8 visually-identical instances, including under partial occlusion by the dispensing head and during liquid transfer.

C.3 Comparative Reasoning Traces

We compare the full reasoning traces of O-VAD against baseline VLMs on two representative success cases from LiquidAD and Phys-AD (Fig. 5).

LiquidAD: unintended single-channel drip. In this case, a multi-channel pipette head releases a droplet from only one tip at frames 30–40, while the remaining tips stay unchanged. Qwen3-VL-32B processes the full video and reports normal operation, noting that the movement is smooth, tips are properly seated, and

no spills or bubbles are observed. Its frame-level descriptions confirm correct aspiration and dispensing but miss the isolated single-tip drip entirely. This is because the anomaly is localized to one channel among many, making it invisible to holistic frame-level inspection. O-VAD, by contrast, tracks the pipette head as a single object with per-frame state annotations. At frames 30–40, the state tracker flags a material release event from only one tip (the second from the right). The reasoning chain then compares this against the expected multi-channel dispense pattern and concludes that an isolated drip is inconsistent with a controlled dispense step. It attributes the cause to a possible seal or pressure imbalance in that channel. This case demonstrates how object-centric tracking surfaces fine-grained per-component deviations that frame-level captioning overlooks.

Phys-AD: sticky roller crushing. Here, a robotic gripper manipulates a paper towel roll mounted on a cardboard core. GPT-5 observes the roller rotating and stopping, describes the behavior as consistent with normal controlled manipulation, and reports no anomaly. Its reasoning explicitly states that no deviations can be established against expectations. O-VAD tracks three objects independently: the cardboard core (obj_id 1), a metal gripper component (obj_id 2), and the paper towel roll (obj_id 3). It detects repeated crushing and flattening of the cardboard core at frames 0–10, 60–70, 140–150, and 160–170, along with compression and bending of the paper roll at frames 110–140. It further identifies a surface tear on the roll end at frames 60–70 and 160–170. The reasoning chain synthesizes these multi-object observations into a coherent diagnosis: excessive grip force or misalignment is causing progressive product damage, supported by concurrent deformation flags on the gripper itself (frames 100–150). GPT-5 misses all of these because its caption-level analysis describes only the high-level motion (rotation and stop) without examining per-object physical integrity.

Both cases illustrate the same bottleneck identified in the main paper: baseline VLMs generate fluent and plausible reasoning but lack fine-grained object-level evidence to anchor their judgments. O-VAD’s structured state trajectories provide the missing evidential layer, enabling the reasoning chain to detect subtle, localized anomalies and to correlate changes across multiple objects for causal diagnosis.

C.4 Failure Cases Analysis

We present four representative failure cases that reveal the fundamental limitations of perception-driven anomaly reasoning without predefined expert knowledge. These cases fall into two categories: *specification-dependent anomalies*, where the ground truth is defined by quantitative thresholds invisible to visual inspection, and *perception-ambiguous anomalies*, where the defect produces visually plausible behavior indistinguishable from normality.

Specification-dependent failures. In the servo angle-restricted case (Fig. 6, top), the servo motor rotates to a small angle and returns smoothly. Both Qwen3-VL-32B and O-VAD classify this as normal, since the observed motion is visually

consistent with a correctly functioning servo. The anomaly is a restricted rotation range, which is defined by a quantitative angular specification entirely absent from the visual evidence. Without knowing the *expected* rotation angle, neither state tracking nor commonsense reasoning can flag this deviation. Similarly, in the clip unable-to-press case (Fig. 7, bottom), the pressing tool descends and retracts while the clip remains static. O-VAD detects no state change events for the clip and concludes no anomaly is present. The ground truth is that the clip *should* have been pressed but was not. This requires knowledge of the expected actuation outcome, which is not visually inferable.

Perception-ambiguous failures. In the button stuck case (Fig. 7, top), O-VAD tracks repeated deformation of the metallic pressing tool across multiple frame intervals. It flags this as tool/fixture damage due to excessive force. However, the actual anomaly is the button’s failure to actuate: the pressing tool deforms precisely *because* the button is stuck. Both a stuck button and a normally stiff button produce similar visual signatures of tool compression. O-VAD captures the correct *symptom* (tool deformation) but misattributes the *cause*, illustrating how accurate state tracking can still lead to incorrect classification when the root cause requires understanding the intended function of the pressed object.

In the magnet degaussing case (Fig. 6, bottom), a degaussed magnet detaches from the board during manipulation. Qwen3-VL-32B attributes this to a gripping mechanism failure. O-VAD detects over-clamping deformation and grip instability, identifying the anomaly as crushing and dropping rather than loss of magnetism. Since the magnet visually resembles an ordinary red plastic block, neither method can infer its magnetic properties. The detachment is caused by insufficient magnetic adhesion, not mechanical failure, but this distinction is invisible to visual perception.

These failure cases share a common pattern: the ground-truth anomaly is defined by an invisible physical property (magnetic strength, required actuation force, angular specification) or requires functional knowledge of what the object is supposed to do. Such information cannot be recovered from visual appearance alone. While O-VAD’s object-centric state tracking correctly captures observable symptoms in most cases, it cannot bridge the gap between perceptual evidence and specification-level judgments. This motivates future work on integrating few-shot specification examples or lightweight expert priors into the reasoning pipeline.

C.5 Human Evaluation

Human Evaluation Protocol. To assess the practical utility of anomaly reports beyond automated metrics, we conduct a human evaluation study with $N=5$ domain experts (graduate researchers with industrial vision or robotics experience). Each evaluator independently reviews anomaly reports generated by three methods—Qwen3-VL-32B (direct prompting), URF-ZS-HVAA, and O-VAD—on 10 randomly sampled videos from the Phys-AD test set, spanning 10 object categories. Reports are anonymized and presented in randomized order to prevent

method-identification bias. Each evaluator completes the questionnaire below for every report.

Instructions. Each evaluator should independently review anomaly reports generated by three methods on 10 randomly sampled videos. For each video–report pair, first watch the video, then read reports, and rate each of the following on a 5-point Likert scale: 1 (strongly disagree) to 5 (strongly agree).

Evaluation Questionnaire. The questionnaire (Table 11) comprises 10 questions organized into five parts that progressively assess report quality from factual correctness to practical utility. Part I (Q1–Q4) measures detection correctness at four granularities: binary anomaly/normal classification, semantic anomaly type matching, affected object identification, and temporal localization. Part II (Q5–Q7) evaluates explanation quality along three complementary axes: faithfulness (no hallucinations), completeness (no omissions), and causal coherence (logically sound reasoning). Part III (Q8) assesses actionability, i.e., whether the report alone suffices for an operator to diagnose and respond. Part IV (Q9) captures a holistic quality judgment. Finally, Part V (Q10) elicits a direct pairwise preference ranking across all methods for the same video, yielding a deployment-oriented win rate. All dimensions except Q10 use a 1–5 Likert scale; Q10 produces a rank ordering from best to worst.

Human Evaluation Results We report results from the human evaluation study described in Appendix C.5. Five domain experts independently rated anomaly reports from three methods—URF-ZS-HVAA (Report 1), Qwen3-VL-32B (Report 2), and O-VAD (Report 3)—on 10 Phys-AD videos (7 anomalous, 3 normal) across 9 Likert-scale dimensions (Q1–Q9) and a pairwise preference ranking (Q10). Table 12 summarizes the results.

Overall comparison. O-VAD achieves the highest scores on all nine dimensions and is ranked first by evaluators in 82% of cases (41 out of 50 evaluator–video pairs). Its average detection score (4.42) exceeds Qwen3-VL-32B [1] (2.89) by +1.53 points and URF-ZS-HVAA [24] (1.14) by +3.28 points. The gap is similarly pronounced for explanation quality (+1.11 over Qwen3, +3.32 over URF-ZS) and actionability (+1.32 over Qwen3, +3.16 over URF-ZS). URF-ZS-HVAA consistently scores near the floor (1.0–1.2) across all dimensions, confirming that its chained reasoning framework, despite incorporating temporal detection and spatial localization modules, produces reports that evaluators judge as almost entirely uninformative for industrial anomaly diagnosis. This is consistent with its near-zero recall in the quantitative evaluation (Table 1 and Table 2).

Detection correctness. The most striking result is on binary detection (Q1), where O-VAD scores 4.98 out of 5—near-perfect agreement that it correctly identifies anomalous vs. normal videos. In contrast, URF-ZS-HVAA scores only 1.24, indicating that evaluators find its predictions almost always incorrect. Qwen3-VL-32B scores 2.70, performing better than URF-ZS but still far below O-VAD.

O-VAD also leads on anomaly type identification (Q2: 4.26 vs. 2.60 vs. 1.08) and temporal localization (Q4: 3.96 vs. 2.32 vs. 1.08), though these finer-grained dimensions show more variance, reflecting the inherent difficulty of open-ended type classification and frame-level precision.

Explanation quality and actionability. Among the explanation dimensions, O-VAD’s largest advantage is on completeness (Q6: 4.58 vs. 3.36 vs. 1.00), indicating that evaluators find its state-trajectory-grounded reports cover all relevant state changes without significant omissions. Causal coherence (Q7: 4.30 vs. 3.02 vs. 1.00) shows the second-largest gap, confirming that the structured chain-of-thought reasoning produces more logically sound root-cause analyses. On diagnostic usefulness (Q8), O-VAD scores 4.16, meaning evaluators believe an operator could largely act on the report without re-watching the video, while URF-ZS-HVAA’s score of 1.00 indicates its reports provide no actionable information.

Anomalous vs. normal cases. On the 7 anomalous videos, O-VAD’s detection average (4.40) substantially exceeds Qwen3-VL-32B (3.19) and URF-ZS-HVAA (1.00). On the 3 normal videos, O-VAD achieves a perfect binary correctness score of 5.00, whereas both baselines score below 1.80 on Q1, indicating severe false-positive tendencies. Notably, Qwen3-VL-32B’s binary correctness drops to 1.13 on normal cases (compared to 3.37 on anomalous ones), suggesting it defaults to predicting anomalies. O-VAD’s robustness on normal videos stems from its evidence-grounded reasoning: without tracked state changes that violate process expectations, the chain-of-thought correctly concludes normality.

C.6 LLM-as-Judge

LLM-as-Judge Evaluation Protocol. To enable scalable evaluation beyond the 10-video human study, we design an LLM-as-judge [47] protocol that mirrors the human questionnaire (§C.5) and can be applied to the full test set. We use GPT-4o as the judge model with structured prompts.

LLM-as-Judge Results. To complement the human evaluation and scale assessment to a larger sample, we apply the LLM-as-judge protocol using GPT-4o as the judge model across the same 10 Phys-AD videos. Note that Q4 (temporal localization) is excluded from this evaluation due to the lack of frame-level ground-truth annotations in the judging set. Table 12 reports the results.

Overall comparison. O-VAD is ranked first by the LLM judge in 70% of cases, followed by Qwen3-VL-32B at 30% and URF-ZS-HVAA at 10%. O-VAD achieves the highest overall quality (Q9: 3.50 vs. 2.70 vs. 1.20) and diagnostic usefulness (Q8: 3.40 vs. 2.30 vs. 1.10), consistent with the human evaluation trends. URF-ZS-HVAA again scores near the floor across all dimensions (1.00–1.40), confirming its inability to produce meaningful anomaly reports regardless of the evaluation paradigm.

Detection vs. explanation: a revealing split. The most notable finding is the divergence between detection correctness and explanation quality. O-VAD leads convincingly on detection (avg. 3.20 vs. 2.33 vs. 1.13) and binary correctness in particular (Q1: 4.60 vs. 2.60 vs. 1.40). However, Qwen3-VL-32B achieves higher explanation scores than O-VAD on all three dimensions: faithfulness (4.90 vs. 3.90), completeness (4.00 vs. 3.90), and causal coherence (3.70 vs. 3.60). This reversal, absent in the human evaluation where O-VAD leads on all explanation metrics, reflects a known LLM-judge bias toward fluent, well-structured prose [47]. Qwen3-VL-32B generates polished natural-language narratives even when its detection is incorrect, which the LLM judge rewards on explanation dimensions evaluated independently of correctness. Human experts, by contrast, penalize explanations that describe the wrong conclusion, regardless of fluency. Despite this explanation advantage, Qwen3-VL-32B’s poor detection correctness (Q1: 2.60) and low actionability (Q8: 2.30) result in a lower overall quality score and win rate, indicating that the LLM judge’s holistic assessment still correctly prioritizes factual accuracy.

Anomaly type correctness remains challenging. Both O-VAD and Qwen3-VL-32B tie on type correctness (Q2: 1.80), the lowest-scoring dimension for both methods. This reflects the inherent difficulty of open-ended anomaly type classification: even when detection is correct, precisely matching the ground-truth anomaly taxonomy remains challenging for all methods, a gap that motivates future work on domain-adaptive type reasoning.

Consistency with Human Evaluation. The LLM-as-judge results broadly corroborate the human evaluation findings: O-VAD ranks first overall, URF-ZS-HVAA is consistently the weakest, and O-VAD’s primary advantage lies in detection correctness and actionability. The main discrepancy is on explanation quality, where the LLM judge favors Qwen3-VL-32B’s fluent but often factually incorrect narratives. This highlights a complementary strength of human evaluation: experts naturally integrate correctness into their explanation judgments, producing assessments more aligned with real deployment requirements.

D Algorithm of O-VAD

Algorithm 1 summarizes the complete O-VAD pipeline. Stage 1 discovers objects via VLM prompting and segments them with SAM, Stage 2 tracks each object through transformations and queries the VLM on sampled frame pairs to detect open-ended state changes, and Stage 3 performs a six-step chain-of-thought reasoning pass followed by confidence-tiered visual verification to produce the final anomaly report. All three stages are training-free and require no predefined taxonomy.

Algorithm 1 O-VAD: Object-Centric Video Anomaly Detection

Require: Video $\mathcal{V} = \{I_t\}_{t=1}^T$, optional caption **cap**, optional task context **ctx**
Ensure: Anomaly report $\mathcal{R}$

```

1:
2: // Stage 1: Automated Object Grounding
3: Sample candidate frames  $\{I_{t_k}\}_{k=1}^K$  from  $\mathcal{V}$ 
4: for each candidate frame  $I_{t_k}$  do
5:    $\mathcal{O}_{t_k} \leftarrow \phi_{\text{VLM}}(I_{t_k})$  ▷ object inventory: names, descriptions, spatial cues
6: end for
7:  $\mathcal{O} \leftarrow \bigcup_k \mathcal{O}_{t_k}$  ▷ merge and deduplicate
8: for each object  $o_i \in \mathcal{O}$  do
9:    $\mathcal{M}_1^i \leftarrow \text{SAM}(I_1, o_i)$  ▷ grounded segmentation
10: end for
11:
12: // Stage 2: Object-Centric State Tracking
13: // Spatiotemporal partitioning & tubelet construction [35]
14:  $\mathcal{P} \leftarrow \text{TubeletGraph}(\mathcal{V}, \mathcal{M}_1)$  ▷ partition video into tubelets
15: // State tracking: recover missing tracks
16: for each object  $o_i$  with prompt track  $\mathcal{P}_i$  do
17:    $\mathcal{T}_i \leftarrow \mathcal{P}_i \cup \{C \in \mathcal{P} \mid S_{\text{prox}}(C, \mathcal{P}_i) > \tau_{\text{prox}} \wedge S_{\text{sem}}(C, \mathcal{P}_i) > \tau_{\text{sem}}\}$ 
18: end for
19: // Initial state identification
20: for each object  $o_i$  with track  $\mathcal{T}_i$  do
21:    $\tilde{I}_1^i \leftarrow \text{Highlight}(I_1, \mathcal{M}_1^i)$ 
22:    $(\text{desc}_i, \text{state}_i, \text{mat}_i) \leftarrow \phi_{\text{VLM}}(\tilde{I}_1^i)$  ▷ object name, state, material
23: end for
24: // State change detection and understanding (ours)
25:  $\mathcal{E} \leftarrow \emptyset$ 
26: for each object  $o_i$  with track  $\mathcal{T}_i$  do
27:   for each sampled pair  $(t_k, t_{k+1})$  from  $\mathcal{T}_i$  do
28:      $\tilde{I}_{t_k}^i \leftarrow \text{Highlight}(I_{t_k}, \mathcal{M}_{t_k}^i)$ ;  $\tilde{I}_{t_{k+1}}^i \leftarrow \text{Highlight}(I_{t_{k+1}}, \mathcal{M}_{t_{k+1}}^i)$ 
29:      $(d, \text{ty}, \text{ca}, \text{desc}, \text{sev}) \leftarrow \phi_{\text{VLM}}(\tilde{I}_{t_k}^i, \tilde{I}_{t_{k+1}}^i)$  ▷ open-ended state change query
30:     if  $d = \text{yes}$  then
31:        $\mathcal{E} \leftarrow \mathcal{E} \cup \{(t_k, t_{k+1}, \text{ty}, \text{ca}, \text{desc}, \text{sev}, i)\}$ 
32:     end if
33:   end for
34: end for
35:
36: // Stage 3: State-Aware Anomaly Reasoning
37:
38: // Cognitive anomaly reasoning (single VLM call, Steps 1–6)
39: Sample key frames  $\mathcal{F} = \{I_1, I_{\lfloor T/2 \rfloor}, I_T\}$  from  $\mathcal{V}$ 
40:  $(\text{reasoning}, \mathcal{A}_{\text{raw}}) \leftarrow \phi_{\text{VLM}}(\mathcal{O}, \mathcal{E}, \text{ctx}, \text{cap}, \mathcal{F})$  ▷ Steps 1–6: observe → expect →  
compare → cause → classify → severity
41:
42: // Post visual verification
43: for each anomaly  $a \in \mathcal{A}_{\text{raw}}$  do
44:   if  $c_{\text{orig}}(a) \geq \tau_{\text{hi}}$  then ▷ high-confidence: skip verification
45:     Retain  $a$ 
46:   else if  $c_{\text{orig}}(a) \leq \tau_{\text{lo}}$  then ▷ low-confidence: auto-discard
47:     Discard  $a$ 
48:   else
49:      $(\text{verified}, c_{\text{ver}}) \leftarrow \phi_{\text{VLM}}(\mathcal{F}, \text{cap}, a)$  ▷ check anomaly evidence in caption/frames
50:     if  $\text{verified} = \text{true}$  then
51:        $c_{\text{final}}(a) \leftarrow c_{\text{orig}}(a) \cdot c_{\text{ver}}$ 
52:     else
53:        $c_{\text{final}}(a) \leftarrow c_{\text{orig}}(a) \cdot (1 - c_{\text{ver}})$ 
54:     end if
55:     if  $c_{\text{final}}(a) < \tau_{\text{conf}}$  then ▷ low post-verification confidence
56:       Discard  $a$ 
57:     end if
58:   end if
59: end for
60:  $\mathcal{A} \leftarrow \text{retained anomalies}$ 
61:
62: return Report  $\mathcal{R} = (\mathcal{A}, \mathcal{E}, \text{reasoning}, \text{summary})$ 

```

Table 8: Video- and type-level results on Phys-AD [23] across 22 object categories. “†” denotes methods requiring training data; “*” denotes training-free methods. “BERT” is the type-level BERTScore, reported only for methods that emit free-form anomaly-type descriptions. For each row, the best / 2nd-best across all methods is highlighted: video-level metrics in / , type-level BERT in / ; our best results are additionally in **bold**. Blank cells indicate results not available. The 22 categories are split into two side-by-side blocks; “Avg.” reports the average over all categories.

Cat.	Met.	Trad.		Direct		Wkf.		Ours		Cat.	Met.	Trad.		Direct		Wkf.		Ours	
		MNAD [†]	S3R [†]	Qwen3*	GPT5*	URF*	VERA [†]	O-VAD*	MNAD [†]			S3R [†]	Qwen3*	GPT5*	URF*	VERA [†]	O-VAD*		
Ball	Acc	0.637	0.326	0.504	0.677	0.704	0.477	0.570	Magnet	Acc	0.533	0.767	0.544	0.908	0.589	0.631	0.689		
	P	0.681	0.000	0.695	0.667	0.692	0.667	0.742		P	0.680	0.800	0.771	0.968	0.717	1.000	0.700		
	R	0.856	0.000	0.456	0.800	1.000	0.057	0.544		R	0.567	0.867	0.450	0.857	0.633	0.314	0.933		
	F1	0.759	0.000	0.550	0.727	0.818	0.105	0.628		F1	0.618	0.832	0.568	0.909	0.673	0.478	0.800		
	AUC	0.589	0.155	0.528	0.744	0.461	0.548	0.579		AUC	0.512	0.724	0.592	0.364	0.518	0.892	0.576		
	BERT			0.821		0.642		0.862		BERT			0.781		0.643		0.699		
Button	Acc	0.697	0.203	0.330	0.400	0.580	0.462	0.380	Roll. Bear.	Acc	0.433	0.483	0.429	0.625	0.517	0.536	0.536		
	P	0.807	1.000	0.809	0.375	0.967	0.500	0.781		P	0.462	0.000	0.350	0.857	0.509	0.000	0.500		
	R	0.817	0.004	0.212	0.171	0.492	0.029	0.312		R	0.800	0.000	0.269	0.231	0.967	0.000	0.346		
	F1	0.812	0.008	0.337	0.235	0.652	0.054	0.446		F1	0.585	0.000	0.304	0.364	0.667	0.000	0.409		
	AUC	0.526	0.108	0.506	0.456	0.717	0.469	0.393		AUC	0.437	0.016	0.418	0.451	0.304	0.500	0.650		
	BERT			0.780		0.666		0.861		BERT			0.812		0.626		0.850		
Car	Acc	0.758	0.497	0.332	0.677	0.642	0.462	0.742	Rub. Band	Acc	0.567	0.583	0.767	0.817	0.733	0.500	0.633		
	P	0.760	0.801	0.733	0.659	0.761	0.000	0.748		P	0.567	0.600	0.944	1.000	0.719	0.000	0.611		
	R	0.991	0.438	0.171	0.829	0.762	0.000	0.989		R	0.567	0.500	0.567	0.633	0.767	0.000	0.733		
	F1	0.860	0.566	0.278	0.734	0.761	0.000	0.852		F1	0.567	0.545	0.708	0.775	0.742	0.000	0.667		
	AUC	0.599	0.527	0.492	0.617	0.497	0.383	0.575		AUC	0.561	0.463	0.767	0.279	0.783	0.276	0.721		
	BERT			0.809		0.603		0.737		BERT			0.803		0.639		0.784		
Caster Wheel	Acc	0.400	0.833	0.300	0.785	0.367	0.462	0.583	Screw	Acc	0.622	0.600	0.444	0.626	0.689	0.460	0.600		
	P	0.737	0.973	1.000	0.957	0.818	0.000	0.717		P	0.760	0.929	0.619	0.596	0.682	0.432	0.700		
	R	0.311	0.800	0.067	0.629	0.200	0.000	0.733		R	0.633	0.433	0.433	0.683	1.000	0.059	0.700		
	F1	0.438	0.878	0.125	0.759	0.321	0.000	0.725		F1	0.691	0.591	0.510	0.661	0.811	0.096	0.700		
	AUC	0.468	0.830	0.533	0.543	0.410	0.609	0.301		AUC	0.638	0.620	0.450	0.441	0.298	0.400	0.704		
	BERT			0.754		0.626		0.796		BERT			0.779		0.657		0.784		
Clip	Acc	0.642	0.622	0.461	0.585	0.344	0.154	0.403	Servo	Acc	0.671	0.333	0.362	0.508	0.750	0.462	0.447		
	P	0.817	0.703	0.619	0.565	1.000	0.000	0.634		P	0.773	0.833	0.737	0.527	0.752	0.000	0.726		
	R	0.596	0.750	0.500	0.617	0.000	0.000	0.246		R	0.794	0.139	0.233	0.829	0.994	0.000	0.418		
	F1	0.689	0.726	0.553	0.722	0.333	0.000	0.354		F1	0.784	0.238	0.354	0.644	0.856	0.000	0.530		
	AUC	0.628	0.537	0.442	0.518	0.275	0.180	0.423		AUC	0.479	0.452	0.492	0.330	0.342	0.341	0.495		
	BERT			0.798		0.640		0.876		BERT			0.797		0.638		0.851		
Clock	Acc	0.568	0.532	0.644	0.723	0.356	0.462	0.458	Slide	Acc	0.647	0.320	0.387	0.523	0.200	0.462	0.493		
	P	0.669	0.734	0.717	0.698	0.857	0.000	1.000		P	0.802	1.000	0.833	0.542	0.000	0.000	0.766		
	R	0.696	0.466	0.770	0.857	0.041	0.000	0.155		R	0.742	0.150	0.292	0.743	0.000	0.000	0.536		
	F1	0.682	0.570	0.743	0.769	0.077	0.000	0.269		F1	0.771	0.261	0.432	0.626	0.000	0.000	0.631		
	AUC	0.513	0.529	0.581	0.766	0.483	0.500	0.669		AUC	0.534	0.533	0.529	0.528	0.345	0.225	0.415		
	BERT			0.792		0.597		0.900		BERT			0.803		0.639		0.812		
Fan	Acc	0.642	0.570	0.427	0.569	0.383	0.462	0.771	Sph. Bear.	Acc	0.500	0.867	0.483	0.467	0.633	0.500	0.517		
	P	0.748	0.828	0.846	0.667	0.831	0.000	0.792		P	0.500	0.893	0.471	0.417	0.595	0.000	0.509		
	R	0.787	0.537	0.287	0.490	0.220	0.000	0.940		R	0.900	0.833	0.267	0.167	0.833	0.000	0.900		
	F1	0.767	0.652	0.429	0.500	0.348	0.000	0.860		F1	0.643	0.862	0.340	0.238	0.694	0.000	0.651		
	AUC	0.409	0.602	0.566	0.472	0.523	0.389	0.532		AUC	0.547	0.894	0.483	0.440	0.629	0.114	0.641		
	BERT			0.797		0.630		0.890		BERT			0.826		0.662		0.726		
Gear	Acc	0.740	0.198	0.638	0.739	0.513	0.462	0.819	Sticky Roller	Acc	0.511	0.689	0.511	0.720	0.756	0.529	0.689		
	P	0.801	0.000	0.880	1.000	0.808	0.000	0.797		P	0.633	0.944	0.700	0.674	0.788	0.488	0.691		
	R	0.897	0.000	0.633	0.514	0.514	0.000	0.953		R	0.633	0.567	0.467	0.736	0.867	0.064	0.967		
	F1	0.847	0.000	0.737	0.679	0.628	0.000	0.868		F1	0.633	0.708	0.560	0.726	0.825	0.105	0.806		
	AUC	0.418	0.358	0.644	0.397	0.451	0.512	0.879		AUC	0.517	0.778	0.533	0.523	0.713	0.474	0.811		
	BERT			0.815		0.643		0.855		BERT			0.814		0.623		0.689		
Hinge	Acc	0.683	0.700	0.417	0.339	0.550	0.462	0.617	Toothpaste	Acc	0.244	0.589	0.578	0.908	0.567	0.431	0.656		
	P	0.760	1.000	0.917	0.409	0.821	0.000	0.750		P	0.233	0.700	0.684	0.914	0.583	0.417	0.667		
	R	0.844	0.600	0.244	0.514	0.511	0.000	0.733		R	0.222	0.311	0.289	0.914	0.467	0.143	0.622		
	F1	0.800	0.750	0.386	0.456	0.630	0.000	0.742		F1	0.227	0.431	0.406	0.914	0.519	0.213	0.644		
	AUC	0.497	0.716	0.589	0.516	0.544	0.041	0.526		AUC	0.257	0.535	0.578	0.832	0.528	0.579	0.689		
	BERT			0.807		0.643		0.790		BERT			0.788		0.619		0.870		
Liquid	Acc	0.578	0.756	0.822	0.750	0.644	0.683	0.667	U Disk	Acc	0.594	0.613	0.442	0.462	0.500	0.446	0.512		
	P	0.641	0.828	0.923	0.941	0.938	1.000	0.667		P	0.502	0.599	0.445	0.000	0.000	0.444	0.533		
	R	0.833	0.800	0.800	0.533	0.500	0.367	1.000		R	0.975	0.683	0.475	0.000	0.000	0.114	0.200		
	F1	0.725	0.814	0.857	0.681	0.652	0.337	0.800		F1	0.663	0.638	0.460	0.000	0.000	0.182	0.291		
	AUC	0.453	0.733	0.833	0.512	0.691	0.827	0.776		AUC	0.513	0.585	0.442	0.437	0.404	0.311	0.606		
	BERT			0.817		0.640		0.666		BERT			0.785		0.650		0.906		
Lock	Acc	0.678	0.328	0.367	0.785	0.333	0.462	0.361	Zipper	Acc	0.472	0.706	0.406	0.554	0.350	0.462	0.600		
	P	0.682	0.000	0.542	0.957	0.000	0.000	0.593		P	0.590	0.914	0.710	0.875	0.800	0.000	0.653		
	R	0.967	0.000	0.325	0.629	0.000	0.000	0.133		R	0.683	0.617	0.183	0.200	0.033	0.000	0.835		
	F1	0.800	0.000	0.406	0.759	0.000	0.000	0.218		F1	0.633	0.736	0.291	0.326	0.064	0.000	0.733		
	AUC	0.442	0.323	0.388	0.231	0.029	0.557	0.780		AUC	0.262	0.791	0.517	0.477	0.273	0.224	0.388		
	BERT			0.774		0.628		0.909		BERT			0.800		0.649		0.711		
Avg.	Acc	0.635	0.591	0.454	0.639	0.329	0.470	0.592											
	P	0.713	0.739	0.717	0.690	0.853	0.500	0.724											
	R	0.803	0.645	0.368	0.580	0.053	0.050	0.625											
	F1	0.755	0.689	0.486	0.630	0.100	0.091	0.621											
	AUC	0.495	0.555	0.512	0.503	0.426	0.456	0.584											
	BERT			0.798		0.632		0.803											

Table 9: Results on IPAD [26] across 16 industrial scenarios (S: synthetic, R: real). “†”=trained; “*”=training-free. For each (scenario, level, metric), the best/2nd-best across methods is highlighted.

S01												S02												S03												S04																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
Method	Lv.	Acc	P	R	F1	AUC	Acc	P	R	F1	AUC	Lv.	Acc	P	R	F1	AUC	Acc	P	R	F1	AUC	Lv.	Acc	P	R	F1	AUC	Lv.	Acc	P	R	F1	AUC																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															
MNAD-p [†]	vid	0.524	0.500	0.200	0.286	0.509	0.368	0.300	0.375	0.333	0.375	0.579	1.000	0.200	0.333	0.600	0.368	0.000	0.000	0.000	0.000	0.418	0.645	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.389																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
	frm	0.614	0.857	0.029	0.055	0.431	0.780	0.386	0.176	0.242	0.548	0.739	0.000	0.000	0.000	0.000	0.418	0.645	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.419																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
S3R [†]	vid	0.619	0.562	0.900	0.692	0.509	0.737	0.667	0.750	0.706	0.716	0.421	0.000	0.000	0.000	0.000	0.267	0.421	0.000	0.000	0.000	0.000	0.267	0.421	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.133																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
	frm	0.481	0.423	0.882	0.572	0.467	0.744	0.390	0.509	0.442	0.700	0.605	0.261	0.293	0.276	0.443	0.364	0.356	0.976	0.521	0.325	0.000	0.000	0.481	0.423	0.882	0.572	0.467	0.744	0.390	0.509	0.442	0.700	0.605	0.261	0.293	0.276	0.443	0.364	0.356	0.976	0.521	0.325	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 10: Intermediate outputs of O-VAD across the three benchmarks. For each dataset we report source video properties (duration, FPS) and per-stage intermediate outputs: Stage-1 grounded object count and Stage-2 detected state count, each summarized as min. / avg. / max. over the evaluation set.

Dataset	Video Info		Intermediate Results (min. / avg. / max.)	
	Duration (s)	FPS	Stage 1: Grounding → #Objects	Stage 2: Tracking → #States
Phys-AD	0.98 / 2.10 / 4.00	60	1 / 2.31 / 12	1 / 10.61 / 73
LiquidAD	8.20 / 13.77 / 204.50	30	1 / 5.44 / 16	13 / 104.56 / 416
IPAD	1.04 / 22.40 / 40.00	25	1 / 3.27 / 10	1 / 29.88 / 134

Table 11: Human evaluation questionnaire. All dimensions use a 1–5 Likert scale (1 = strongly disagree, 5 = strongly agree) except Q10 (preference ranking).

Part	ID	Dimension & Description	Scale
Det.	Q1	Binary correctness. Correctly identifies anomalous vs. normal?	1–5
	Q2	Type correctness. Predicted type matches the ground-truth defect? [†]	1–5
	Q3	Object identification. Affected object(s) correctly identified? [‡]	1–5
	Q4	Temporal localization. Anomaly accurately localized in time? [‡]	1–5
Expl.	Q5	Faithfulness. Explanation reflects actual video content, free of hallucinations.	1–5
	Q6	Completeness. All salient state changes covered without significant omissions.	1–5
	Q7	Causal coherence. Causal reasoning is logically sound and evidence-consistent.	1–5
Act.	Q8	Diagnostic usefulness. Operator could act on the report without re-watching the video.	1–5
Ovr.	Q9	Overall quality. Overall quality of the anomaly report.	1–5
Pref.	Q10	Preference. Rank all method reports for the same video for real deployment.	Rank 1–3

[†] Scored only when both ground truth and prediction are anomalous. [‡] Scored only when provided.



Fig. 5: Comparative reasoning traces on LiquidAD and Phys-AD. Top: An unintended single-channel drip during automated pipetting. Qwen3-VL-32B observes the full sequence but concludes normal operation, while O-VAD tracks per-tip state changes and identifies the isolated droplet release at frames 30–40. **Bottom:** A sticky roller undergoes crushing during robotic manipulation. GPT-5 describes the rotation as normal and finds no deviation, while O-VAD tracks progressive deformation of the cardboard core and paper roll across multiple frame intervals, detecting crushing, surface tearing, and possible end-effector overload.

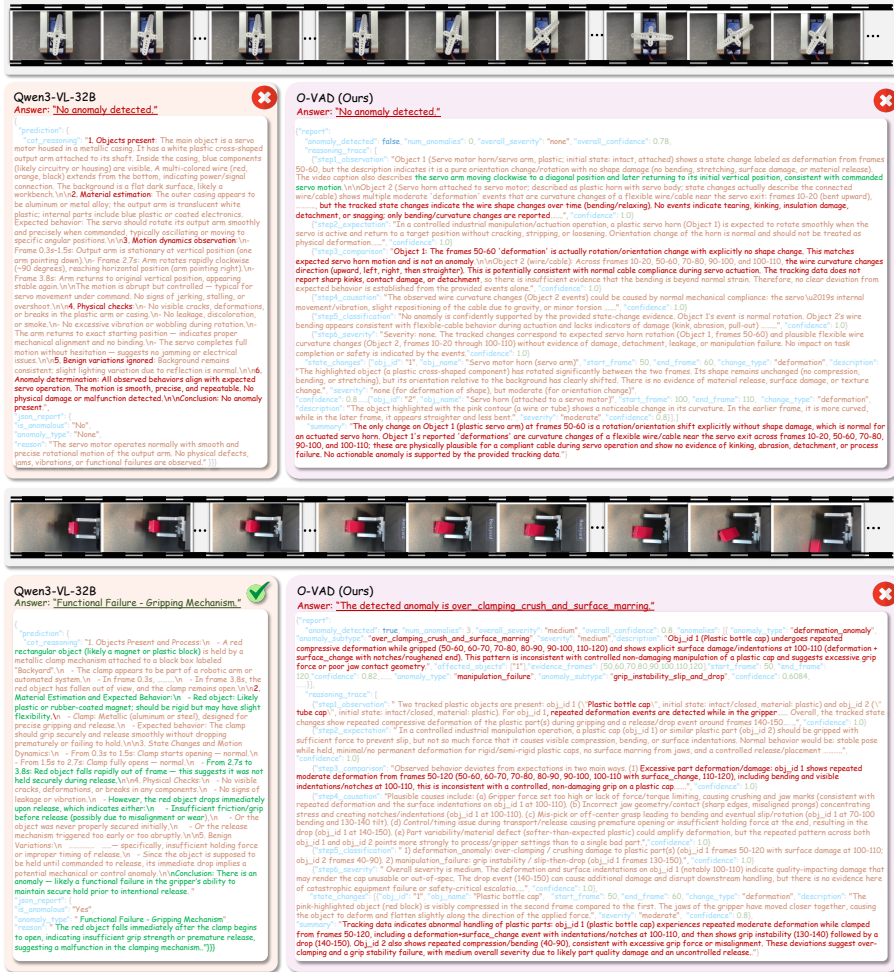


Fig. 6: Failure cases on servo and magnet categories. *Top:* A servo motor with a restricted rotation angle is misclassified as normal by both Qwen3-VL-32B and O-VAD. The observed small-angle motion is visually indistinguishable from a normal commanded rotation. *Bottom:* A degaussed magnet falls off the board. Qwen3-VL-32B incorrectly attributes the anomaly to a gripping mechanism failure. O-VAD detects over-clamping deformation but fails to identify the root cause as loss of magnetism, since the magnet visually resembles an ordinary red plastic block.

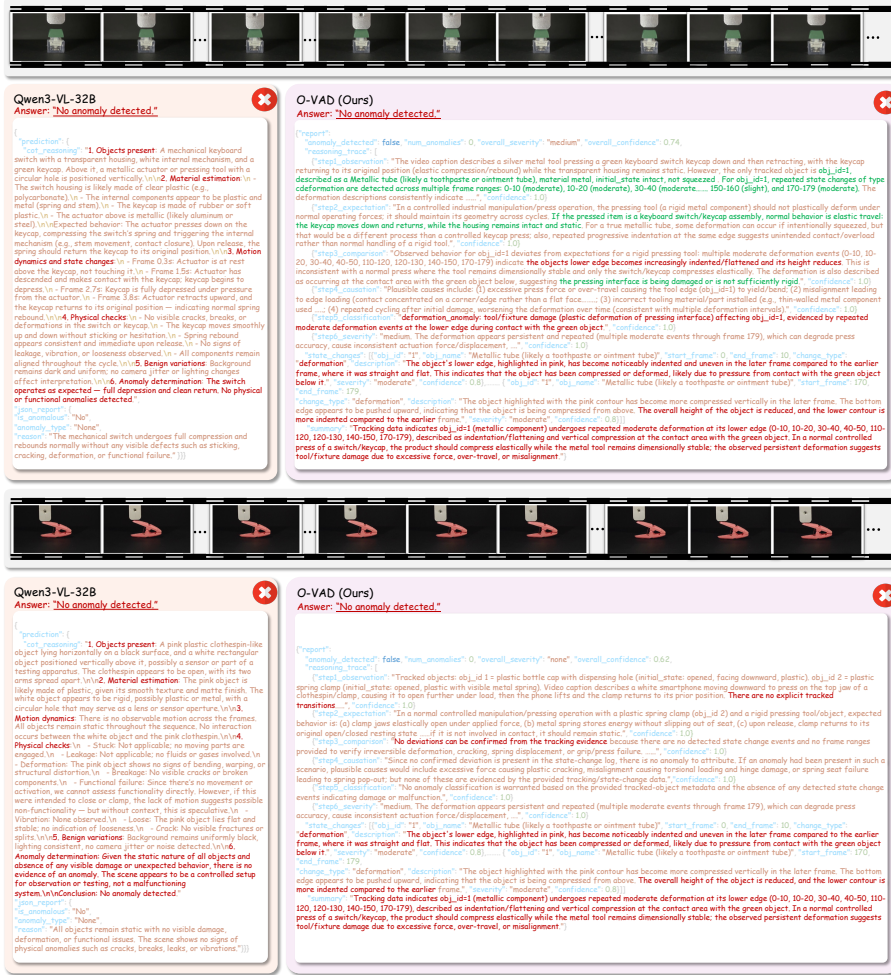


Fig. 7: Failure cases on button and clip categories. *Top:* A stuck button that fails to actuate is misclassified as normal by both methods. O-VAD detects repeated deformation of the pressing tool but attributes it to tool/fixture damage rather than recognizing the button's inability to depress. *Bottom:* A clip that cannot be pressed shows no observable state changes, leading both methods to predict normal behavior despite the ground-truth anomaly.

Table 12: Human and LLM-as-judge evaluation results (mean scores, 1–5 scale). Human: $N=5$ evaluators $\times$ 10 videos; LLM: GPT-4o judge $\times$ 10 videos. URF-ZS: URF-ZS-HVAA [24]; Qwen3: Qwen3-VL-32B [1]. Q4 is excluded from LLM evaluation due to the absence of temporal ground truth. Best per row per evaluator is **bold**; ties share bold.

Part	Dimension	Human Evaluation			LLM-as-Judge		
		URF-ZS	Qwen3	O-VAD	URF-ZS	Qwen3	O-VAD
Det.	Q1. Binary correctness	1.24	2.70	4.98	1.40	2.60	4.60
	Q2. Type correctness	1.08	2.60	4.26	1.00	1.80	1.80
	Q3. Object identification	1.14	3.94	4.46	1.00	2.60	3.20
	Q4. Temporal localization	1.08	2.32	3.96		<i>N/A</i>	
	<i>Detection avg.</i>	1.14	2.89	4.42	1.13	2.33	3.20
Expl.	Q5. Faithfulness	1.04	3.28	4.10	1.30	4.90	3.90
	Q6. Completeness	1.00	3.36	4.58	1.20	4.00	3.90
	Q7. Causal coherence	1.00	3.02	4.30	1.20	3.70	3.60
	<i>Explanation avg.</i>	1.01	3.22	4.33	1.23	4.20	3.80
Act.	Q8. Diagnostic usefulness	1.00	2.84	4.16	1.10	2.30	3.40
Ovr.	Q9. Overall quality	1.04	3.00	4.28	1.20	2.70	3.50
Pref.	Q10. Win rate (%)	0.0	18.0	82.0	10.0	30.0	70.0