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

Mei Yuan[1], Qi Long[1], Qifeng Wu[1], Zhenyang Li[2], Yizhou Zhao[1], Lei Wang[3], Yang Liu[1], and Min Xu[1⋆]

[1] Carnegie Mellon University, USA
[2] University of Alabama at Birmingham, USA
[3] Griffith University, Australia
meiyuan@andrew.cmu.edu, mxu1@cs.cmu.edu
https://o-vad.github.io/

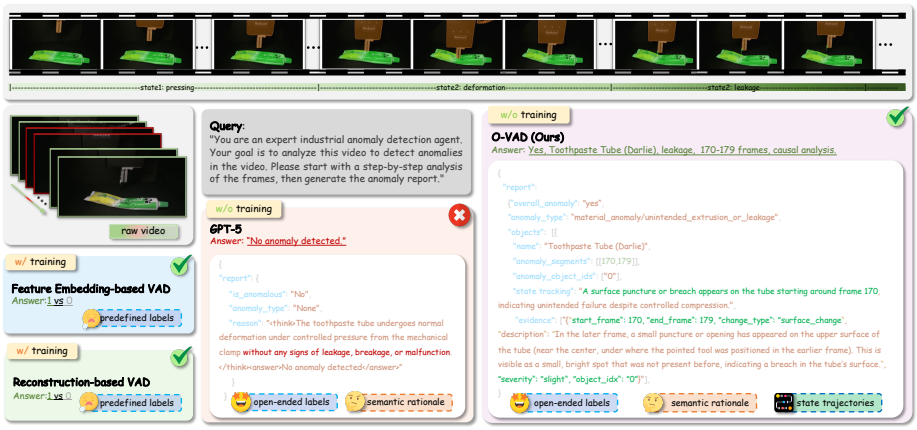

Fig. 1: Why object state evolution matters for industrial anomaly detection. The toothpaste tube undergoes pressing and deformation before a subtle leakage emerges. Traditional methods predict binary classifications. VLM prompting fails to detect it due to the lack of evidence. O-VAD tracks object-wise 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 [22]. 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 [5, 6, 9, 36–38] reconstruct anomalous samples to their corresponding normal counterparts and calculate the reconstruction error, and embedding-based [4,8,12,14,28] 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 [7], 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 [7,11,15,20,21,39,41,42] 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 [13] 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 [15]. (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 [16,23,30,33,34], and detect anomalies from when and where perspectives [10, 20]. 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 withy 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 [5, 6,9,36–38] flag anomalies by their elevated reconstruction error, while embedding-based methods [4, 8, 12, 14, 28] 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 [7], requiring abundant normal samples and a separate model per object class, and must be re-trained for every unseen domain or anomaly type [18], making them impractical for novel categories and dynamic production lines. Second, they are largely black-box, emitting an anomaly score without semantic rationale [34], which has motivated growing interest in explainable, language-grounded anomaly understanding [18, 40]. Third, they compress entire frames into holistic features, neglecting region-level cues and over-relying on dominant background context [17]. 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 [7, 11, 15, 20, 21, 39, 41, 42] 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 [16, 23, 30, 33, 34]; for instance, a unified zero-shot framework [20] 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

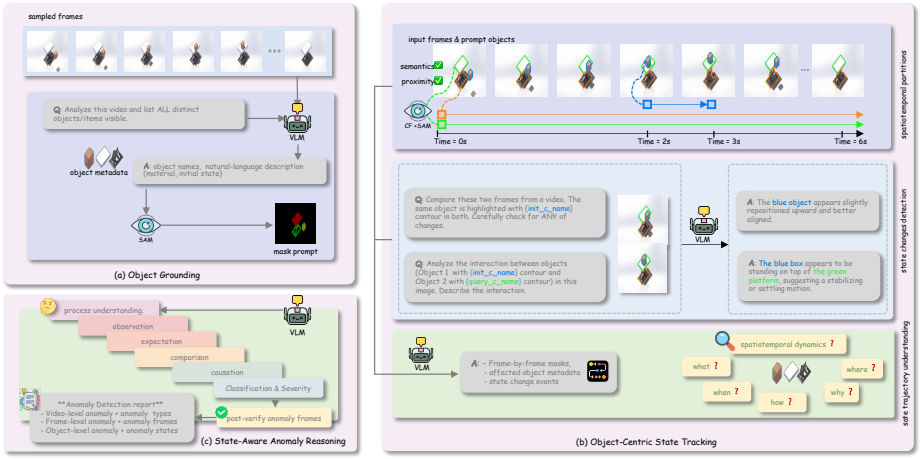

Fig. 2: Overview of O-VAD. (a) Stage 1 samples multiple frames and queries a VLM to discover all objects, then segments them with SAM to produce initial masks and metadata. **(b)** Stage 2 constructs spatiotemporal tubelets via CropFormer [25] and SAM2 [27], recovers missing tracks through semantic and 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:

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

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:

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

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]:

$$m_k^{(1)} = \mathrm{SAM}(I_1,\, o_k), \qquad \mathcal{M}_1 = \big\{m_k^{(1)}\big\}_{k=1}^{|\mathcal{O}|}. \tag{3}$$

The output of Stage 1, including initial masks $\mathcal{M}_1$ together with object metadata, 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 [31] 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 [25]) produces per-frame spatial partitions $\mathcal{E}_t = \mathrm{CF}(I_t)$, and each entity $e_1^i \in \mathcal{E}_1$ (including the initial object masks $\mathcal{M}_1$) is tracked forward via SAM2 [27] to form tubelets:

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

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 $C = \{c_s, c_{s+1}, \ldots, c_T\}$ beginning at frame s and the prompt object track $P = \{p_1, p_2, \ldots, p_T\}$, the spatial proximity score is:

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

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 [26] feature $f(M, I) = \text{Pool}(\text{CLIP}(I), M)$ via mask-pooling. The semantic similarity is:

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

The final set of valid continuation tracks is:

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

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:

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

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, $\textit{sev} \in \{\text{none}, \text{slight}, \text{moderate}, \text{severe}\}$ 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 miscategorized.
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:

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

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:

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

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* [19]: 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* [3]: 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* [22]: industrial-manufacturing processes (∼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 (Sec. A.1 and Sec. 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 [29] 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 [25] for per-frame entity segmentation and SAM2 [27] 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 (Sec. A.2 and Sec. 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 [24] (unsupervised, prediction-based memory normality) and S3R [32] (weakly-supervised

Table 1: Video-level, type-level, and frame-level Acc/P/R/F1/AUROC (%) on Phys-AD [19], LiquidAD [3], and IPAD [22]. 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.

			[Trad.]		[Direct Prompting]		[Workflow]		[O-VAD(Ours)]
Dataset	**Level**	**Metric**	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	video	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**
	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 [29]—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 [20] chains temporal detection, spatial localization, and textual explanation via intra- and inter-task reasoning; VERA [35] 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 (↑) on Phys-AD [19] across 22 object categories. "†" = methods requiring training data. Per category, the best / 2nd-best across all methods is / ; our best results are additionally in **bold**.

	Trad.		Direct Prompting		Workflow		O-VAD (Ours)
Category	MNAD.p†	S3R†	Qwen3*	GPT-5*	URF*	VERA†	**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 → track → 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 (Sec. B.1 and Sec. B.2).

4.3 Qualitative Results

To illustrate how object-centric evidence shapes reasoning quality, we compare O-VAD against GPT-5 [29]—prompted with the same cascaded chain-of-thought but *without* object grounding or state tracking—on three representative Phys-AD scenarios (Figure 3). Additional reasoning traces, failure-case analyses, and human and LLM-as-judge studies are provided in the appendix (Sec. C).

Case Analysis. Figure 3 visualizes full reasoning traces on representative Phys-AD 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.

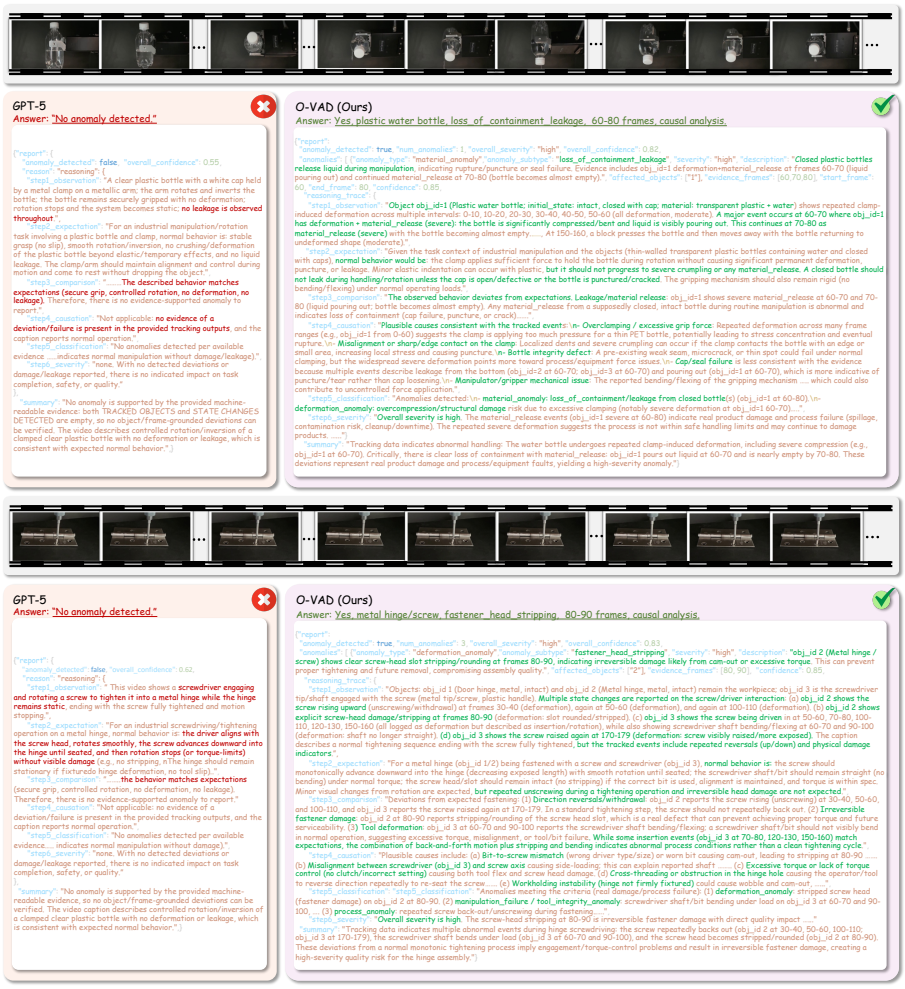

Fig. 3: Qualitative results. Each row shows a different industrial manipulation task—plastic bottle rotation (top), hinge screw fastening (bottom)—with GPT-5's output (left) and O-VAD's output (right).

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
O-VAD (full)	**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) *w/o Caption*	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) *w/o State Track.*	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) *w/o CoT Reas.*	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) *w/o Post-verifier*	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
O-VAD (full)	**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) *w/o Caption*	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) *w/o State Track.*	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) *w/o CoT Reas.*	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) *w/o Post-verifier*	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 commonsense 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. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. 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)
12. Hyun, J., Kim, S., Jeon, G., Kim, S.H., Bae, K., Kang, B.J.: Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2052–2061 (2024)
13. 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)
14. Jiang, X., Liu, J., Wang, J., Nie, Q., Wu, K., Liu, Y., Wang, C., Zheng, F.: Softpatch: Unsupervised anomaly detection with noisy data. Advances in Neural Information Processing Systems **35**, 15433–15445 (2022)
15. 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)

16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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)
26. 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)
27. 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)
28. 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)
29. 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)
30. 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)
31. Sun, Y., Yang, X., Sun, J.J., Hariharan, B.: Tracking and understanding object transformations. arXiv preprint arXiv:2511.04678 (2025)

32. 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)
33. 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)
34. 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)
35. 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)
36. Zavrtanik, 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)
37. Zhang, H., Wang, Z., Wu, Z., Jiang, Y.: Diffusionad: norm-guided one-step denoising diffusion for anomaly detection (2023)
38. 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)
39. 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)
40. 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)
41. Zhu, L., Chen, Q., Shen, X., Cun, X.: Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. arXiv preprint arXiv:2505.23504 (2025)
42. 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)