ADCLAW: EVIDENCE-GROUNDED ACTIVE INSPECTION FOR INDUSTRIAL ANOMALY DETECTION AND REASONING

Anonymous authors
Paper under double-blind review

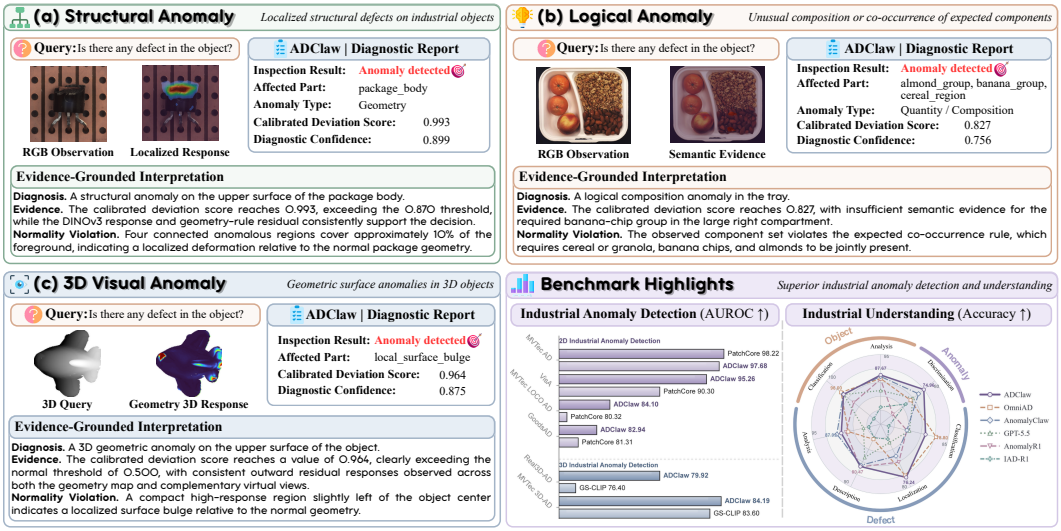


Figure 1: **Representative results of ADClaw.** Evidence-grounded diagnoses of structural, logical, and 3D anomalies, together with comparative results on 2D/3D IAD benchmarks and MMAD.

ABSTRACT

Industrial anomaly detection is essential for automated industrial inspection. Although multimodal large language models offer strong visual and semantic reasoning capabilities, existing methods often rely on a single observation or use tool outputs as transient feedback. This limits their ability to retain evidence across inspection steps and connect intermediate findings to final diagnoses. We introduce **ADClaw**, an evidence-grounded agent that formulates industrial anomaly diagnosis as active evidence construction and hypothesis verification. Its **Structured Normality Prior** integrates object–part semantics, explicit normality rules, and multi-scale normal visual memory to identify inspection targets and select suitable normal references. Guided by this prior, a persistent, programmable **Inspection Workbench** organizes heterogeneous tool outputs into a structured evidence state, allowing intermediate findings to be traced, composed, and reused throughout inspection. **Hypothesis-Guided Evidence Reasoning** assesses the sufficiency and consistency of this accumulated evidence to decide whether to acquire further observations, revise the current hypothesis, or finalize the diagnosis. These components work together in a *hypothesize–inspect–reason–revise* loop, turning discrete tool calls into a stateful inspection process in which each action addresses an explicit evidence requirement. Evaluations on 2D and 3D industrial anomaly detection benchmarks and MMAD demonstrate competitive detection and localization performance while enabling evidence-grounded defect understanding and traceable diagnostic explanations linking intermediate findings to final decisions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

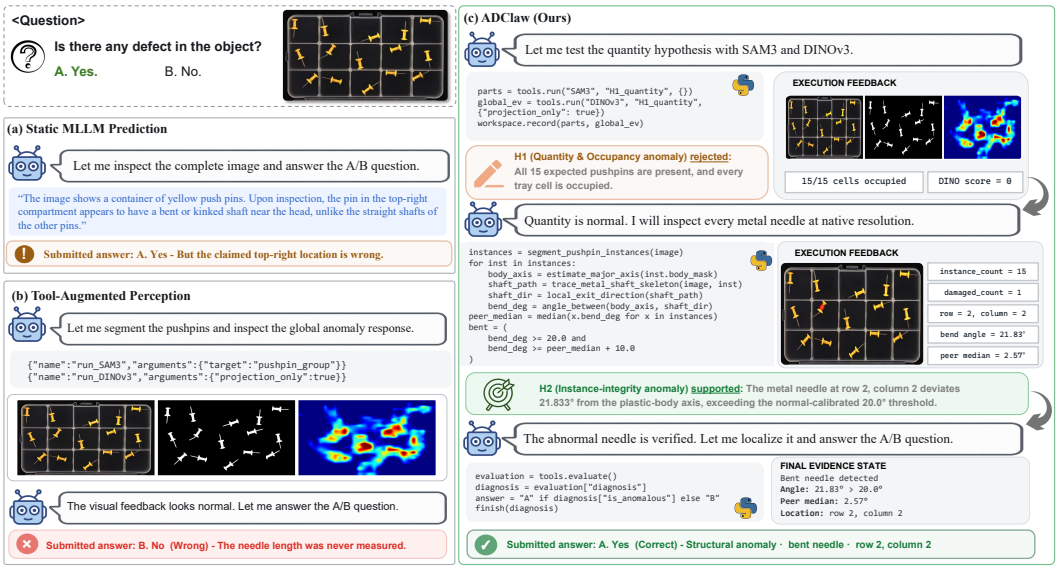


Figure 2: **Qualitative comparison on MVTec LOCO.** Static MLLM prediction (a) and tool-augmented perception with one-time feedback (b) fail to provide reliable localized evidence, whereas ADClaw (c) maintains an evolving evidence state for hypothesis-guided verification.

1 Introduction

Industrial anomaly detection (IAD) aims to identify and localize deviations from normal conditions in industrial products and is a key visual task in intelligent manufacturing (Bergmann et al., 2019; Zou et al., 2022). Recent advances in multimodal large language models (MLLMs) have introduced stronger visual understanding and reasoning capabilities to IAD, enabling models to localize anomalies, identify defect types, and generate natural language explanations (Gu et al., 2024; Li et al., 2026c; Xu et al., 2025; Jiang et al., 2025). However, industrial anomalies are often *small-scale, low-contrast, highly localized*, and *structure-dependent* (Heckler-Kram et al., 2025; Bergmann et al., 2022a). Reliable diagnosis therefore requires close examination of local details, relationships between components, and subtle deviations from normal conditions; patch-level normal-feature matching, local vision–language alignment, and component-aware comparison provide complementary ways to examine these cues (Roth et al., 2022; Jeong et al., 2023; Gu et al., 2025). We therefore reformulate IAD from a static image prediction task into an **executable inspection process**: an agent actively acquires, combines, and verifies visual evidence around its current *anomaly hypothesis*, using this evidence to determine *what to inspect, how to inspect, and when to revise the hypothesis*.

Several MLLM-based IAD methods improve anomaly perception and answer generation through anomaly-aware visual representations or task-specific post-training (Gu et al., 2024; Xu et al., 2025; Chao et al., 2025; Li et al., 2026b). These advances motivate a complementary question: how should heterogeneous intermediate evidence be represented and reused throughout an extended inspection process? Recent agentic approaches augment visual reasoning with local observation, normal-reference retrieval, and anomaly-specific tools, enabling multi-round inspection with accumulated interaction history or reference memory (Miao et al., 2026; Jiang et al., 2026; Zhang et al., 2026b). For industrial inspection, evidence organization is consequential because a diagnosis may depend on both component relationships and explicit normality constraints (Gu et al., 2025; Zhang et al., 2025b). For example, the agent may need to identify a target component, inspect a local region, and compare appearance or geometric evidence against normal references, drawing on complementary information from different representations (Roth et al., 2022; Wang et al., 2023). Programmatic visual systems and executable code actions provide mechanisms for composing tools and processing their outputs (Gupta & Kembhavi, 2023; Surís et al., 2023; Wang et al., 2024). Building on these capabilities, our focus is to organize heterogeneous observations into a unified, programmable evidence state and explicitly assess its sufficiency and consistency for the current anomaly hypothesis. Figure 2 illustrates this distinction on the same industrial inspection sample,

comparing static MLLM prediction, tool-augmented perception with transient visual feedback, and ADClaw's iterative evidence-driven inspection process.

To realize this executable inspection paradigm, we introduce the **Inspection Workbench**, a persistent computational workspace that integrates tool orchestration, evidence organization, and state updates. Rather than treating tool outputs as isolated, transient feedback, the workbench incorporates them into an *inspection state* whose contents can be continually updated, combined, and reused. This allows the agent to iteratively acquire and examine diagnostic evidence around its current anomaly hypothesis and adjust its inspection strategy accordingly. Thus, the workbench composes tool outputs into reusable evidence across inspection steps.

To address the limited object-specific normality constraints provided by generic visual knowledge and the difficulty of maintaining consistency between evidence from multiple sources and the final diagnosis, we construct an **Industrial Semantic Knowledge Base**. It uses object–part structure to associate normality rules with defect semantics and maps these priors to the corresponding inspection tools, providing industrial knowledge to guide evidence acquisition. In parallel, a **Normal Visual Memory** organizes multi-scale visual prototypes derived from normal samples into a semantic–topological hierarchy. Conditioned on the current anomaly hypothesis, the agent routes retrieval requests to memory nodes that match the inspection target and scale, providing relevant normal references for examining both overall structure and local details. Building on these components, we introduce **Hypothesis-Guided Evidence Reasoning**, which combines retrieved normality priors with accumulated inspection evidence to assess consistency among anomaly decisions, spatial localization, defect types, and their explanations. When the available evidence is insufficient or conflicting, the agent replans its inspection actions and revises the current hypothesis. This establishes a *hypothesize–inspect–reason–revise* loop for addressing surface, structural, and logical anomalies within a unified framework. As illustrated in Fig. 1, ADClaw connects anomaly decisions with supporting visual evidence and normality constraints across both 2D and 3D inspection scenario. Our contributions are as follows:

- **Evidence-Grounded Inspection Workbench.** We introduce a persistent, programmable workbench that composes heterogeneous tools and maintains reusable evidence states for computation across inspection steps.
- **Hypothesis-Guided Evidence Reasoning.** We evaluate evidence sufficiency and consistency to guide further inspection, hypothesis revision, or diagnosis, establishing an adaptive loop between evidence acquisition and diagnostic reasoning.
- **Systematic Evaluation across Industrial Benchmarks.** We evaluate ADClaw on 2D and 3D benchmarks covering surface, structural, and logical anomalies, assessing detection, localization, anomaly understanding, and the contributions of individual components.

2 Related Work

Industrial Anomaly Detection. Industrial anomaly detection has developed several approaches to normality modeling, including reconstruction learning, knowledge distillation, and feature memory banks (Zavrtanik et al., 2021; Deng & Li, 2022; Roth et al., 2022). With advances in foundation models, recent studies have explored vision–language alignment, prompt learning, and local feature matching with pretrained models to improve detection in zero-shot and few-shot settings (Jeong et al., 2023; Zhou et al., 2024a; Cao et al., 2024; Damm et al., 2025). Modeling part relationships, compositional matching, and appearance–geometry fusion have further extended detection to structural, logical, and multimodal anomalies (Gu et al., 2025; Zhang et al., 2025a; Wang et al., 2023). These methods provide useful evidence for anomaly detection and localization. Our focus is on organizing such evidence into explicit, verifiable explanations of observed defects. Following PatchCore, we match DINOv3 descriptors against normal features to produce anomaly maps for evidence-grounded diagnosis (Roth et al., 2022; Siméoni et al., 2025).

Multimodal Anomaly Understanding. Multimodal large language models have extended industrial anomaly analysis beyond scoring and localization to defect description, semantic explanation, and interactive question answering. AnomalyGPT uses simulated anomalies and corresponding textual descriptions for supervision (Gu et al., 2024). Myriad incorporates priors from visual experts, while Anomaly-OV emphasizes anomaly-relevant visual tokens (Li et al., 2026c; Xu et al., 2025).

Reinforcement learning with structured rewards has also been explored to improve anomaly perception and reasoning (Chao et al., 2025; Li et al., 2026b). Meanwhile, MMAD provides a systematic evaluation of anomaly discrimination, defect understanding, and object understanding (Jiang et al., 2025). Complementing model adaptation and post-training, ADClaw organizes evidence at inference time while keeping its pretrained model parameters fixed.

Tool-Augmented Visual Agents. Tool-augmented visual agents integrate observation, external execution, and feedback into multi-step reasoning processes. VISPROG and ViperGPT compose visual modules through generated programs (Gupta & Kembhavi, 2023; Surís et al., 2023). CodeAct reuses Python variables across code actions, while SpatialClaw employs a stateful kernel for spatial reasoning (Wang et al., 2024; Cho et al., 2026). In industrial anomaly detection, program-driven inspection has been used to check logical constraints (Zhang et al., 2025b). AgentIAD combines supervised fine-tuning and reinforcement learning to learn policies for multi-round observation and reference retrieval (Miao et al., 2026). Recent training-free methods further support diagnosis through normal references, multi-round critique, anomaly-specific tools, and reference memory (Jiang et al., 2026; Zhang et al., 2026b). ADClaw extends this direction with a persistent Inspection Workbench for evidence reuse and composition. Object-specific priors and hypothesis-guided action selection determine whether to reanalyze evidence, inspect another region, or finalize its diagnosis.

3 METHODS

3.1 OVERVIEW

ADClaw performs industrial anomaly diagnosis through iterative, evidence-grounded inspection, as illustrated in Fig. 3. The **Structured Normality Prior** (SNP) (Section 3.2) links object–part knowledge and normality rules to executable tools and normal visual references, specifying inspection targets and comparison criteria. Guided by this prior, the planner organizes targeted checks through the **Inspection Workbench** (IW) (Section 3.3), which retains intermediate tool outputs as reusable evidence across inspection steps. The inspection process supports hypothesis refinement in 2D and primarily complementary virtual-view selection in 3D. Finally, **Hypothesis-Guided Evidence Reasoning** (HER) (Section 3.4) assesses the accumulated evidence to guide further inspection and form the final diagnosis. Diagnostic outputs are fixed before evidence-grounded explanation generation.

3.2 STRUCTURED NORMALITY PRIOR

For category c, the Structured Normality Prior combines semantic constraints and normal visual references:

$$\mathcal{P}_c = (\mathcal{K}_c, \mathcal{M}_c), \qquad \mathcal{K}_c = (\mathcal{O}_c, \mathcal{R}_c, \mathcal{T}), \tag{1}$$

where $\mathcal{O}_c$, $\mathcal{R}_c$, and $\mathcal{T}$ denote the object–part ontology, normality rules, and tool contracts, respectively. The visual memory $\mathcal{M}_c$ is constructed from the normal reference set $\mathcal{N}_c$.

Semantic knowledge. The ontology $\mathcal{O}_c$ specifies part hierarchies, observable attributes, and grounding prompts; normality rules $\mathcal{R}_c$ define conditions on component counts, geometric integrity, and spatial relations, together with required evidence and checks. Tool contracts $\mathcal{T}$ specify operation inputs, outputs, and applicability. Shared object and part identifiers map inspection targets to their constraints, prompts, and executable checks.

Normal visual memory. Using frozen encoders, ADClaw builds a *semantic-topological hierarchical memory* $\mathcal{M}_c$ from normal-sample features, with object-level global representations, semantically aligned part representations, and local visual patterns. It retains sample features and cluster prototypes indexed by category, part, and sample. Part–feature associations encode semantics, while local cluster layouts and adjacency statistics capture normal 2D organization. For each target, knowledge-guided hierarchical retrieval (KHR) filters references by global similarity, part match, and local layout; adjacency statistics flag atypical component arrangements.

For 2D nearest-neighbor comparison, let $\hat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$ denote a normalized query descriptor and $\mathcal{B}$ the corresponding set of normalized normal features. The feature deviation is

$$d(\mathbf{z}, \mathcal{B}) = \mathrm{clip}_{[0,1]}\left(1 - \max_{\mathbf{b}\in\mathcal{B}} \hat{\mathbf{z}}^\top \mathbf{b}\right), \tag{2}$$

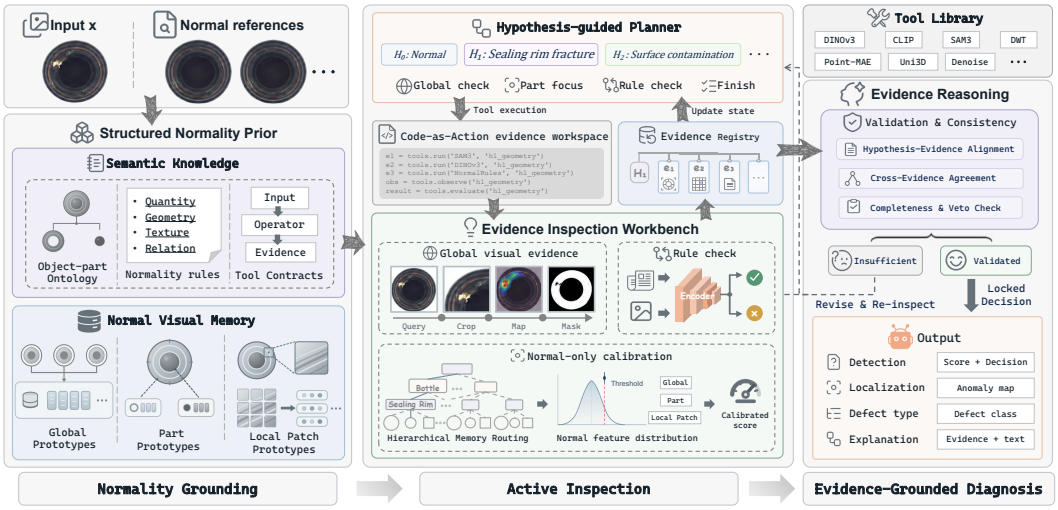


Figure 3: **Overview of ADClaw.** A structured normality prior combines semantic knowledge with normal visual memory to guide hypothesis-driven inspection. The planner executes tool-based checks in a persistent workbench, where intermediate results are retained as reusable, traceable evidence. Hypothesis-Guided Evidence Reasoning (shown as Evidence Reasoning) evaluates whether the accumulated evidence supports the current hypotheses and guides further inspection or hypothesis revision when necessary. The diagnosis is fixed before explanation generation.

where $\mathrm{clip}_{[0,1]}$ clips values to $[0, 1]$. Global and part deviations use their corresponding prototypes. Target-conditioned patch comparisons use retrieved normal features, whereas dense responses use the full category-specific normal patch memory.

For 3D inspection, memory stores normal point-cloud templates and frozen point-cloud features for geometric comparison after registration and feature matching, respectively. Normal-sample statistics calibrate the scoring branches. Appendix C details memory configurations and tool implementations.

3.3 Inspection Workbench

The Inspection Workbench maintains a persistent evidence state. Given input x, normality prior $\mathcal{P}_c$, and accumulated evidence $\mathcal{E}_t$, the planner selects action a_t to address the current hypothesis's evidence requirements. The workbench executes the action and updates the state:

$$\mathcal{E}_{t+1} = \mathcal{E}_t \cup \mathrm{Execute}(a_t; x, \mathcal{P}_c, \mathcal{E}_t). \tag{3}$$

Tool outputs are registered in the Evidence Registry, with their key contents summarized as

$$e = \big(\mathrm{id}_e, h_e, r_e, \tau_e, s_e^{\mathrm{raw}}, \widetilde{s}_e, q_e, \Omega_e, \mathrm{provenance}_e\big). \tag{4}$$

Here, id_e is a unique identifier; h_e and r_e denote the associated hypothesis and applicable rules; and τ_e identifies the tool. The raw score s_e^{raw}, calibrated score $\widetilde{s}_e$, and quality coefficient q_e summarize the result. Ω_e gives its spatial scope, and $\mathrm{provenance}_e$ records its source and upstream evidence IDs. Rule associations reside in hypothesis or tool metadata. Inapplicable numerical fields may remain undefined; localization and semantic outputs remain structured evidence. Masks, feature responses, and normal references are reusable, with dependencies linking derived measurements to source evidence.

For 2D inspection, the shared full-image response A is restricted to target region Ω_h, yielding $A|_{\Omega_h}$; composition, structural, and semantic evidence support hypothesis verification and refinement. For 3D, geometric evidence guides complementary virtual-view selection to examine anomalous regions. Hypothesis-Guided Evidence Reasoning evaluates the evidence and returns unresolved requirements to the planner.

3.4 HYPOTHESIS-GUIDED EVIDENCE REASONING

Hypothesis-Guided Evidence Reasoning (HER) tests whether accumulated evidence supports candidate hypothesis h. ADClaw retrieves the latest eligible scored record from each applicable tool into $\mathcal{E}_h$, then selects decision evidence $\mathcal{D}_h \subseteq \mathcal{E}_h$. Rather than averaging tool outputs, it computes a quality-aware hypothesis score:

$$s_h = \frac{\sum_{e \in \mathcal{D}_h} w_e q_e \widetilde{s}_e}{\sum_{e \in \mathcal{D}_h} w_e q_e}, \tag{5}$$

where $\widetilde{s}_e$ is the calibrated anomaly score, q_e its quality, and w_e the predefined tool weight. Unscored artifacts add no numerical support. Auxiliary evidence may verify rules, provide semantic support, or trigger a conditional veto, but is not fused unconditionally.

ADClaw further evaluates *cross-evidence consistency* in score, spatial, and semantic dimensions:

$$C_h = \mathbf{1}[\sigma_h \le \tau_{\text{score}}]\, \mathbf{1}[\iota_h \ge \tau_{\text{loc}}]\, \mathbf{1}[\rho_h \ge \tau_{\text{sem}}]\,. \tag{6}$$

Here, σ_h, ι_h, and ρ_h measure score dispersion, bounding-box agreement, and semantic-label agreement (see Appendix C.3 for their computation and applicability). A hypothesis is accepted only with sufficient evidence, a score above its threshold, and satisfied consistency, spatial, support, and veto conditions:

$$A_h = \mathbf{1}[|\mathcal{E}_h| \ge m]\, \mathbf{1}[s_h > \tau_h]\, C_h G_h R_h (1 - V_h). \tag{7}$$

Here, G_h, R_h, and V_h denote the spatial check, additional support requirement, and veto condition. Thus, a high score alone cannot validate a hypothesis.

If evidence is incomplete or inconsistent, its gaps return to the planner for further inspection or hypothesis revision within the budget. After validation, the anomaly score, decision, and localization are fixed before explanation generation; the evidence-grounded explanation does not alter these diagnostic outputs.

Table 1: **2D anomaly detection performance on MVTec AD, VisA, MVTec LOCO AD, and GoodsAD**. I-R and P-R denote image- and pixel-level AUROC; F1 denotes F1-max. All metrics are reported as percentages (%). **Bold** and underline mark the best and second-best results.

Category	Method	MVTec AD			VisA			MVTec LOCO AD			GoodsAD		
		I-R↑	P-R↑	F1↑	I-R↑	P-R↑	F1↑	I-R↑	P-R↑	F1↑	I-R↑	P-R↑	F1↑
Traditional IAD	PatchCore (2022)	**98.22**	97.96	97.81	90.30	98.27	87.54	80.32	**88.27**	81.07	81.31	**97.78**	79.25
	FastFlow (2021)	91.46	95.64	93.16	87.90	96.03	84.90	75.46	69.71	80.29	60.69	75.40	73.34
VLM-based IAD	WinCLIP (2023)	70.11	80.97	86.14	64.26	70.05	74.23	53.47	52.87	77.36	54.59	79.81	71.24
	AnomalyCLIP (2024a)	91.55	91.07	92.75	81.97	95.47	80.44	61.86	63.75	77.33	57.25	83.49	71.39
	AnomalyR1 (2025)	83.66	–	89.50	77.37	–	78.64	58.32	–	77.49	58.32	–	71.50
	Qwen3.5-9B (2026)	95.69	–	95.41	87.92	–	84.28	71.11	–	79.23	53.53	–	71.80
	ADClaw	97.68	**98.22**	**98.33**	**95.26**	**98.99**	**92.57**	**84.10**	85.73	**83.22**	**82.94**	97.01	**81.35**

4 EXPERIMENTS

Experimental setup. We evaluate 2D anomaly detection on MVTec AD, VisA, MVTec LOCO AD, and GoodsAD, 3D detection on Real3D-AD and MVTec 3D-AD, and anomaly understanding on MMAD (38 categories). We report image-level AUROC and F1-max, and pixel-level AUROC for 2D detection; object-level AUROC and average precision, and point-level AUROC for 3D detection. MVTec 3D-AD additionally reports P-PRO, the normalized area under the PRO–FPR curve up to 0.30 FPR. Detection and localization metrics are macro-averaged over categories. Our MMAD evaluations report seven task accuracies and their mean, using balanced accuracy for anomaly discrimination. Appendix A details the protocols.

Baselines and implementation. We compare ADClaw with traditional and VLM-based IAD methods, 3D detectors, and anomaly-understanding models. ADClaw uses Qwen3.5-9B with all pretrained parameters frozen; normal references support visual memory and calibration. The 2D toolkit combines DINOv3 ViT-L/16, SAM3, CLIP, and the discrete wavelet transform (DWT). DINOv3 uses 448×448 inputs and features from layers 6, 12, and 18. For 3D inspection, Geometry3D and frozen PointMAE features provide geometric and feature evidence, with at most six 256×256 virtual views per scan. Appendices B–C detail baselines, memory construction, and tool settings.

Table 2: **Comparison of 3D industrial anomaly detection performance on Real3D-AD and MVTec 3D-AD.** O-R and O-AP denote object-level AUROC and average precision; P-R denotes point-level AUROC. P-PRO is the normalized area under the PRO–FPR curve up to 0.30 FPR, reported only on MVTec 3D-AD. **Bold** and underline mark the best and second-best model scores.

Category	Method	Real3D-AD			MVTec 3D-AD			
		O-R↑	O-AP↑	P-R↑	O-R↑	O-AP↑	P-R↑	P-PRO↑
Traditional 3D IAD	BTF (2023)	60.10	61.00	57.11	60.55	85.45	77.41	49.25
	PatchCore-FPFH (2022)	58.88	56.97	62.48	80.29	93.62	**96.80**	89.02
	PatchCore-PointMAE (2022)	58.53	60.41	64.81	70.03	87.97	75.26	65.49
	M3DM-PointBERT (2023)	53.76	58.38	61.82	69.99	89.34	95.07	89.43
	M3DM-PointMAE (2023)	53.82	55.41	62.28	68.25	88.41	95.28	**90.12**
Language-guided 3D IAD	BTP (2026a)	61.40	65.10	**84.50**	–	–	–	–
	PointAD (2024b)	74.80	76.90	73.50	82.00	94.20	95.50	84.40
	GS-CLIP (2026)	76.40	77.70	76.30	83.60	**96.50**	96.30	86.40
Ours	**ADClaw**	**79.92**	**81.16**	83.59	**84.19**	94.19	92.82	82.49

4.1 2D Industrial Anomaly Detection

Table 1 shows a consistent image-level benefit across different 2D inspection scenarios: ADClaw leads on seven of eight image-level metrics, with its largest I-AUROC gains over the Qwen baseline occurring on MVTec LOCO AD and GoodsAD. This pattern is consistent with the value of normality-grounded inspection for compositional constraints and product-specific appearance variations. Normal references, object–part semantics, and explicit rules jointly contextualize the visual discrepancies observed during inspection. On these two datasets, ADClaw improves image-level discrimination despite PatchCore retaining stronger pixel-level localization, suggesting that localizing discrepancies and judging overall abnormality make different demands on the available evidence. Overall, the results support combining visual comparison with structured inspection to address diverse industrial anomalies while keeping pretrained model parameters fixed.

4.2 3D Industrial Anomaly Detection

The 3D results in Table 2 support the applicability of evidence-grounded inspection to both unorganized point clouds and organized scans. ADClaw attains the highest O-AUROC on both benchmarks and balances strong object-level detection with competitive point-level localization on Real3D-AD. Here, registration residuals and normal-feature comparisons provide different forms of numerical evidence, which guide tool selection and complementary views of the same scan. On MVTec 3D-AD, object-level gains do not imply leading point-level localization, reflecting the distinct demands

Table 3: **Comparison of anomaly understanding performance on MMAD (accuracy, %). Bold** and underline mark the best and second-best model scores.

Method	Scale	Anomaly	Defect				Object		Average
		Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Random Chance (2025)	–	50.00	25.00	25.00	25.00	25.00	25.00	25.00	28.57
Human (ordinary) (2025)	–	86.90	66.25	85.58	71.25	81.52	89.58	69.72	78.69
Human (expert) (2025)	–	95.24	75.00	92.31	83.33	94.20	86.11	80.37	86.65
GPT-4o (2024)	–	68.63	65.80	55.62	73.21	83.41	94.98	82.80	74.92
GPT-5.5 (2026)	–	69.95	64.91	62.34	75.95	80.10	95.23	82.11	76.20
Gemini-1.5-Pro (2024)	–	68.63	60.12	58.56	70.38	82.46	89.20	82.25	73.09
Claude-3.5-Sonnet (2024)	–	60.14	58.34	48.81	67.13	79.11	85.19	79.83	68.36
Qwen3.5-9B (2026)	9B	65.03	62.64	63.53	76.82	84.79	92.87	86.93	76.09
LLaVA-NeXT (2024)	34B	57.92	48.79	52.87	71.34	80.28	81.12	77.80	67.16
InternVL2 (2024)	76B	68.25	54.22	56.66	66.30	80.47	86.40	82.92	70.75
ADSeeker (2026a)	7B	68.62	53.82	55.57	60.51	77.40	89.91	83.48	69.90
IAD-R1 (2026b)	7B	66.62	54.15	60.95	66.53	78.55	86.90	76.11	69.97
AnomalyR1 (2025)	3B	60.62	63.56	70.14	**80.47**	85.28	92.48	86.15	76.96
AnomalyClaw (2026)	–	72.51	63.64	67.73	79.65	**87.99**	95.38	87.13	79.15
OmniAD (2026)	7B	68.80	**78.80**	75.50	67.20	86.40	**96.00**	86.40	79.90
ADClaw	9B	**74.96**	73.31	**76.24**	80.27	86.68	95.29	**87.67**	**82.06**

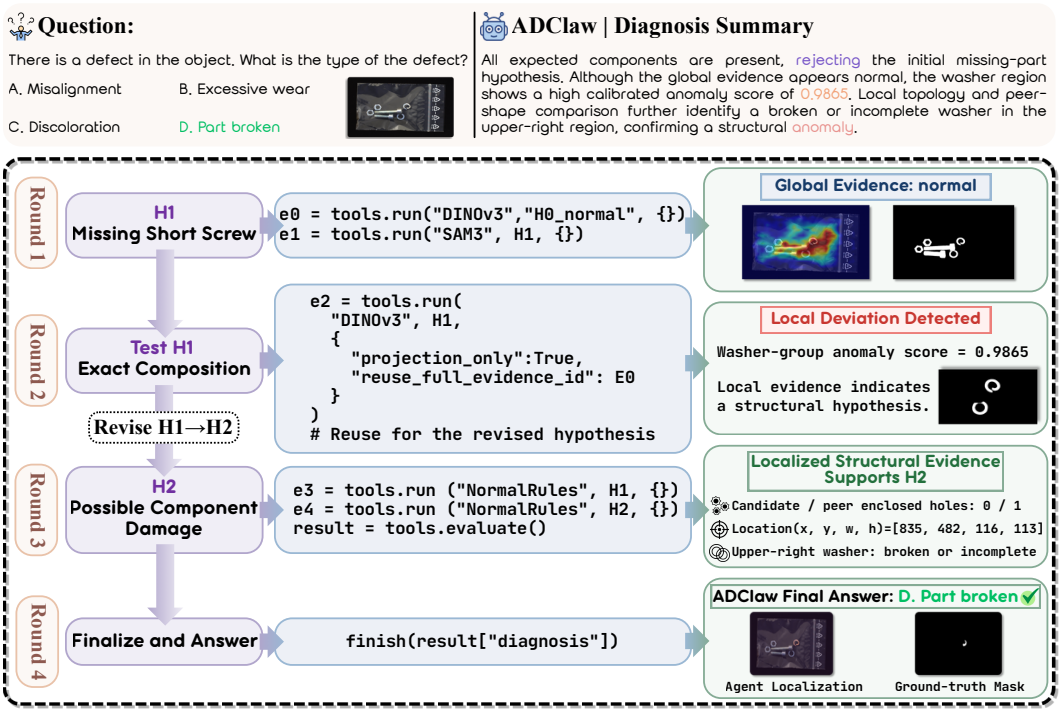


Figure 4: **A summarized inspection trace for an MMAD defect-classification example.** ADClaw rejects an initial missing-screw hypothesis and identifies a broken washer through part-level visual evidence, composition checks, and component-integrity evaluation.

of identifying an anomalous object and delineating its defective surface. Taken together, the results are consistent with using a common evidence-centered inspection process across 3D representations, supported by dataset-specific geometric and feature scoring branches.

4.3 Anomaly Understanding

The distribution of gains in Table 3 suggests that ADClaw primarily improves defect-specific judgment rather than general object recognition. Relative to the same 9B Qwen backbone, the largest gains occur in defect classification and localization, at 10.67 and 12.71 percentage points, respectively, whereas gains on the two object-understanding tasks are smaller. This is consistent with the role of active evidence construction: normal references and targeted measurements constrain both the defect category and its spatial location, giving reasoning a concrete basis for revising an initial interpretation. Figure 4 illustrates this process: accumulated part-level evidence rejects a missing-screw hypothesis and supports a broken-washer diagnosis. ADClaw's top-three ranking across all seven tasks further supports pairing visual evidence with hypothesis verification.

4.4 Ablation Study

Contribution of key mechanisms. Table 4 evaluates KHR within the Structured Normality Prior, the Inspection Workbench (IW), and Hypothesis-Guided Evidence Reasoning (HER). Removing any mechanism reduces overall accuracy, with larger effects on anomaly discrimination, defect classification, and localization than on object semantics. Removing KHR reduces defect localization from 76.24% to 70.40%, supporting target-matched normal references for distinguishing defects from appearance variation. Removing IW lowers localization to 71.80%, indicating the value of retaining intermediate evidence across inspection steps. Removing HER produces the largest discrimination drop (74.96% to 70.80%), highlighting the value of sufficient, consistent evidence.

Effect of stateful evidence-driven inspection. Table 5 isolates inspection strategy while holding the backbone, tools, normal references, preprocessing, and evaluation protocol fixed. *Fixed Tool Pipeline* executes a predefined sequence of operations. *Adaptive Tool Calling* selects tools from previous observations but treats their outputs as transient feedback. ADClaw additionally retains

Table 4: **Ablation of key mechanisms on full MMAD (accuracy, %).** KHR denotes knowledge-guided hierarchical retrieval within the normality prior; IW, the persistent Inspection Workbench; and HER, Hypothesis-Guided Evidence Reasoning.

Variant	Anomaly	Defect				Object		Average
	Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Qwen3.5-9B	65.03	62.64	63.53	76.82	84.79	92.87	86.93	76.09
w/o KHR	71.80	68.90	70.40	78.60	85.50	94.30	87.10	79.51
w/o IW	72.90	70.10	71.80	78.90	85.80	94.60	87.20	80.19
w/o HER	70.80	69.50	73.00	79.10	85.20	94.80	87.00	79.91
ADClaw (Full)	**74.96**	**73.31**	**76.24**	**80.27**	**86.68**	**95.29**	**87.67**	**82.06**

structured evidence for reuse, hypothesis revision, and consistency validation. Adaptive selection improves the MMAD average from 78.73% to 80.32%; retaining and validating evidence further raises it to 82.06%. The latter gains are largest for anomaly discrimination and defect localization, supporting the contribution of evidence organization beyond access to the same perception tools.

Table 5: **Inspection strategies on full MMAD (accuracy, %).** Variants share the backbone, visual tools, normal references, preprocessing, and evaluation protocol.

Inspection Strategy	Anomaly	Defect				Object		Average
	Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Fixed Tool Pipeline	69.80	67.95	69.88	78.55	85.00	94.30	85.60	78.73
Adaptive Tool Calling	72.40	70.85	72.65	79.30	85.80	94.75	86.50	80.32
ADClaw (Full)	**74.96**	**73.31**	**76.24**	**80.27**	**86.68**	**95.29**	**87.67**	**82.06**

Efficiency of reusable evidence. We measure evidence reuse through paired, fixed-trace replay on 100 MVTec LOCO images (20 per category). Both settings execute the same recorded actions and differ only in reusing persisted full-image DINO responses for part-level projections. In Figure 5, reuse reduces mean DINO forward passes from 3.85 to 1.00 per image (**74.0%**) and tool-stage replay latency from 5.61 to 4.19 seconds (**25.4%**). Models and normal references are preloaded; timing excludes Qwen planning and subsequent postprocessing. All 100 paired runs pass output-equivalence checks, preserving the recorded actions and outputs. The workbench reduces repeated visual computation while preserving evidence for subsequent inspection.

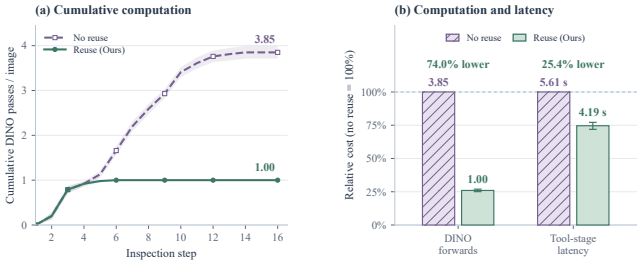


Figure 5: **Efficiency of evidence reuse.** (a) Cumulative DINO passes. (b) Computation and tool-stage latency relative to no reuse. Raw means are annotated; bands and bars show image-level bootstrap 95% confidence intervals.

5 Conclusion

We presented ADClaw as an evidence-grounded framework for industrial anomaly detection and reasoning, and evaluated it across 2D, 3D, and anomaly-understanding benchmarks. The results show that organizing and validating intermediate evidence can improve defect-oriented reasoning while maintaining competitive detection and localization performance, and that persistent evidence reuse can also reduce redundant visual computation during multi-step inspection. These findings highlight stateful evidence management as a useful design direction for tool-augmented industrial inspection and more reliable multi-stage diagnostic reasoning, especially when decisions depend on heterogeneous and spatially distributed evidence. Future work will focus on more robust evidence acquisition for difficult local defects, stronger cross-modal consistency checking, and broader generalization to unseen industrial categories and inspection settings.

AI USE STATEMENT

During manuscript preparation, we used AI tools to assist with writing and language polishing, literature retrieval, and figure and table design and formatting. Writing assistance was used to refine the description of ADClaw and the appendix, improve clarity and terminological consistency, and support the interpretation of experimental results. Literature retrieval assistance helped identify relevant studies on industrial anomaly detection, anomaly understanding, and tool-augmented reasoning. It also supported the organization of knowledge-base materials and entry descriptions, including the organization and presentation of object ontologies, normality rules, and tool contracts. Figure and formatting assistance was used for framework diagrams, inspection traces, qualitative examples, and LaTeX tables. We have manually checked all AI-assisted content and revised it where necessary, and we take full responsibility for the content of the final manuscript.

REPRODUCIBILITY STATEMENT

We will open-source the implementation of ADClaw to support reproduction and further research. Appendix A describes the datasets, normal-reference usage, and evaluation metrics. Appendix B provides the baseline implementation settings and dataset-specific configurations. Appendix C details the structured normality priors, normal visual memory, tool contracts, and inspection procedures for 2D and 3D inputs. Appendix D reports category-level detection and localization results on the 2D and 3D benchmarks. ADClaw keeps pretrained model parameters fixed, using normal-reference samples for memory construction and calibration rather than additional training or fine-tuning. These descriptions, together with the planned code release, will support reproduction of the method and evaluation procedures.

REFERENCES

Anthropic. Claude 3.5 Sonnet, June 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Bergmann_MVTec_AD_--_A_Comprehensive_Real-World_Dataset_for_Unsupervised_Anomaly_CVPR_2019_paper.html.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization. *International Journal of Computer Vision*, 130(4):947–969, February 2022a. ISSN 1573-1405. doi: 10.1007/s11263-022-01578-9. URL http://dx.doi.org/10.1007/s11263-022-01578-9.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pp. 202–213, 2022b. doi: 10.5220/0010865000003124. URL https://arxiv.org/abs/2112.09045.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/05096.pdf.

Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. AnomalyR1: A GRPO-based end-to-end MLLM for industrial anomaly detection. *arXiv preprint arXiv:2504.11914*, 2025. URL https://arxiv.org/abs/2504.11914.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, and Min-Hung

Chen. SpatialClaw: Rethinking action interface for agentic spatial reasoning. *arXiv preprint arXiv:2606.13673*, 2026. URL https://arxiv.org/abs/2606.13673.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Damm_AnomalyDINO_Boosting_Patch-Based_Few-Shot_Anomaly_Detection_with_DINOv2_WACV_2025_paper.html.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9737–9746, 2022. URL https://arxiv.org/abs/2201.10703.

Zehao Deng, An Liu, and Yan Wang. Gs-clip: Zero-shot 3d anomaly detection by geometry-aware prompt and synergistic view representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35587–35596, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Deng_GS-CLIP_Zero-shot_3D_Anomaly_Detection_by_Geometry-Aware_Prompt_and_Synergistic_CVPR_2026_paper.html.

Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i3.27963. URL https://ojs.aaai.org/index.php/AAAI/article/view/27963.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. UniVAD: A training-free unified model for few-shot visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.03342.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2211.11559.

Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection, 2025. URL https://arxiv.org/abs/2503.21622.

Eliahu Horwitz and Yedid Hoshen. Back to the feature: Classical 3d features are (almost) all you need for 3d anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2968–2977, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/VAND/html/Horwitz_Back_to_the_Feature_Classical_3D_Features_Are_Almost_All_CVPRW_2023_paper.html.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Jeong_WinCLIP_Zero-Few-Shot_Anomaly_Classification_and_Segmentation_CVPR_2023_paper.html.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=JDiER86r8v.

Xi Jiang, Yinjie Zhao, Zesheng Yang, and Feng Zheng. AnomalyClaw: A universal visual anomaly detection agent via tool-grounded refutation. *arXiv preprint arXiv:2605.10397*, 2026. URL https://arxiv.org/abs/2605.10397.

Kaiqiang Li, Gang Li, Mingle Zhou, Min Li, Delong Han, and Jin Wan. Back to point: Exploring point-language models for zero-shot 3d anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14167–14177, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Back_to_Point_Exploring_Point-Language_Models_for_Zero-Shot_3D_Anomaly_CVPR_2026_paper.html.

Yanhui Li, Yunkang Cao, Chengliang Liu, Yuan Xiong, Xinghui Dong, and Chao Huang. IAD-R1: Reinforcing consistent reasoning in industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6583–6591, 2026b. doi: 10.1609/aaai.v40i8.37588. URL https://ojs.aaai.org/index.php/AAAI/article/view/37588.

Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. Myriad: A large multimodal model applying vision experts for industrial anomaly detection. *Science China Information Sciences*, 69(9):192103, 2026c. doi: 10.1007/s11432-025-4976-0. URL https://link.springer.com/article/10.1007/s11432-025-4976-0.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jiaqi Liu, Guoyang Xie, Ruitao Chen, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, and Feng Zheng. Real3D-AD: A dataset of point cloud anomaly detection. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/611b896d447df43c898062358df4c114-Abstract-Datasets_and_Benchmarks.html.

Junwen Miao, Penghui Du, Yingying Fan, Yi Liu, Yu Wang, Runze He, Lida Huang, and Yan Wang. AgentIAD: Agentic industrial anomaly detection via adaptive memory augmentation. *arXiv preprint arXiv:2512.13671*, 2026. URL https://arxiv.org/abs/2512.13671v2. Version 2, revised April 14, 2026; first submitted in 2025.

OpenAI. GPT-4o system card, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-5.5 system card, April 2026. URL https://openai.com/index/gpt-5-5-system-card/.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, July 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Roth_Towards_Total_Recall_in_Industrial_Anomaly_Detection_CVPR_2022_paper.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Suris_ViperGPT_Visual_Inference_via_Python_Execution_for_Reasoning_ICCV_2023_paper.html.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50208–50232. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24h.html.

Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2303.00601.

Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M. Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2502.07601.

Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows, 2021. URL https://arxiv.org/abs/2111.07677.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM – a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zavrtanik_DRAEM_-_A_Discriminatively_Trained_Reconstruction_Embedding_for_Surface_Anomaly_ICCV_2021_paper.html.

Jian Zhang, Runwei Ding, Miaoju Ban, and Linhui Dai. PKU-GoodsAD: A supermarket goods dataset for unsupervised anomaly detection and segmentation. *IEEE Robotics and Automation Letters*, 9(3):2008–2015, 2024. doi: 10.1109/LRA.2024.3352358. URL https://doi.org/10.1109/LRA.2024.3352358.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. URL https://arxiv.org/abs/2503.18325.

Kai Zhang, Zekai Zhang, Xihe Sun, Anpeng Wang, Jingmeng Nie, Qinghui Chen, Han Hao, Jianyuan Guo, and Jinglin Zhang. Adseeker: A knowledge-grounded reasoning framework for industry anomaly detection and reasoning, 2026a. URL https://arxiv.org/abs/2508.03088v2. Version 2, revised April 3, 2026; first submitted in 2025.

Yi Zhang, Jiawen Zhu, Lele Fu, and Guansong Pang. AnomalyAgent: Training-free agentic models for zero-/few-shot anomaly detection. *arXiv preprint arXiv:2605.30140*, 2026b. URL https://arxiv.org/abs/2605.30140.

Yiheng Zhang, Yunkang Cao, Xiaohao Xu, and Weiming Shen. LogiCode: An LLM-driven framework for logical anomaly detection. *IEEE Transactions on Automation Science and Engineering*, 22:7712–7723, 2025b. doi: 10.1109/TASE.2024.3468464. URL https://ieeexplore.ieee.org/document/10710633/.

Shifang Zhao, Yiheng Lin, Yunpeng Huan, Weizhe Liu, Lu Han, Yao Zhao, and Yunchao Wei. OmniAD: Intrinsic anomaly detection and reasoning via semantic anomaly encoding, 2026. URL https://arxiv.org/abs/2505.22039v2. Version 2, revised July 29, 2026; first submitted in 2025.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=buC4E91xZE.

Qihang Zhou, Jiangtao Yan, Shibo He, Wenchao Meng, and Jiming Chen. Pointad: Comprehending 3d anomalies from points and pixels for zero-shot 3d anomaly detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84866–84896. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2695. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9a263e235f6d1521d13a8531c7974951-Paper-Conference.pdf.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference Self-supervised Pre-training for Anomaly Detection and Segmentation. In *Computer Vision – ECCV 2022*, pp. 392–408. Springer Nature Switzerland, 2022. ISBN 9783031200564. doi: 10.1007/978-3-031-20056-4_23. URL http://dx.doi.org/10.1007/978-3-031-20056-4_23.

SUPPLEMENTARY MATERIAL

This appendix provides evaluation protocols (Section A), baseline settings (Section B), implementation details (Section C), additional results (Section D) and qualitative analyses (Section E).

A DATASETS AND EVALUATION PROTOCOLS

A.1 DATA ORGANIZATION AND NORMAL REFERENCES

We evaluate ADClaw on benchmarks covering different forms of industrial anomalies. MVTec AD (Bergmann et al., 2019) includes object and texture defects, VisA (Zou et al., 2022) contains single- and multi-instance objects, MVTec LOCO AD (Bergmann et al., 2022a) evaluates both structural and logical anomalies, and GoodsAD (Zhang et al., 2024) covers diverse retail products. For 3D evaluation, Real3D-AD (Liu et al., 2023) provides unorganized point clouds, whereas MVTec 3D-AD (Bergmann et al., 2022b) preserves an organized XYZ grid. Both 3D configurations use XYZ observations. Table A1 summarizes the evaluated datasets and their normal-reference splits.

Table A1: Dataset statistics and available normal-reference samples. Normal train and normal val. denote the sizes of the normal training and validation splits, respectively; the test columns report the full test splits. Normal references support memory construction and score calibration.

Dataset	Classes	Normal train	Normal val.	Test normal	Test anomalous
MVTec AD	15	3,629	–	467	1,258
VisA	12	8,659	–	962	1,200
MVTec LOCO AD	5	1,778	305	575	993
GoodsAD	6	3,137	–	1,328	1,659
Real3D-AD	12	48	–	604	602
MVTec 3D-AD	10	2,656	294	249	948

Use of normal references. ADClaw requires no additional training or fine-tuning of its pretrained models. Instead, normal samples provide the visual references and statistics against which test observations are compared. Memory construction and per-query reference selection serve distinct purposes: the former establishes a reusable comparison basis, while the latter supplies references relevant to the current inspection. For example, the VisA memory contains features from 8,659 normal images, although the evaluation adapter initially supplies one normal reference per query. GoodsAD retrieves one or two matched product views with their normal-image descriptions.

For 3D inspection, Real3D-AD uses four normal templates per category and obtains calibration statistics from normal-template comparisons. In MVTec 3D-AD, each query is associated with four normal references, while its PointMAE memory is constructed from 96 normal training scans per category. The corresponding scoring branches use normal validation scans for calibration. Reference construction and calibration do not update pretrained model parameters.

A.2 EVALUATION METRICS

Image- and object-level anomaly detection. We use image-level AUROC (I-AUROC) for 2D anomaly detection and object-level AUROC (O-AUROC) for 3D anomaly detection. Both metrics measure the ability of continuous anomaly scores to distinguish anomalous samples from normal ones. Anomalous samples are treated as the positive class, and a sample is predicted as anomalous when its score is greater than or equal to a threshold θ. Varying this threshold yields the true-positive rate and false-positive rate:

$$\begin{aligned} \mathrm{TPR}(\theta) &= \frac{\mathrm{TP}(\theta)}{\mathrm{TP}(\theta) + \mathrm{FN}(\theta)}, \\ \mathrm{FPR}(\theta) &= \frac{\mathrm{FP}(\theta)}{\mathrm{FP}(\theta) + \mathrm{TN}(\theta)}. \end{aligned} \tag{A1}$$

Here, TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively. AUROC is the area under the resulting receiver operating characteristic curve and summarizes discrimination performance across thresholds.

For 3D detection, we additionally report object-level average precision (O-AP), which evaluates precision across recall levels. We use non-interpolated average precision:

$$\text{O-AP} = \sum_{j=1}^{J} (r_j - r_{j-1})\, p_j, \tag{A2}$$

where J is the number of evaluation points corresponding to distinct score thresholds, and p_j and r_j are the precision and recall at the j-th point. The points are ordered by increasing recall, with $r_0 = 0$. Precision is the proportion of predicted anomalous samples that are truly anomalous, whereas recall is the proportion of anomalous samples correctly detected.

For 2D detection, we also report image-level F1-max, defined as the maximum F1 score obtained by varying the decision threshold:

$$\text{F1}_{\max} = \max_{\theta} \frac{2\,\text{TP}(\theta)}{2\,\text{TP}(\theta) + \text{FP}(\theta) + \text{FN}(\theta)}. \tag{A3}$$

Samples with identical scores are processed together during threshold sweeping. F1-max measures the best F1 attainable on the evaluated scores, not the F1 at the system's fixed decision threshold.

Pixel- and point-level anomaly localization. P-AUROC denotes pixel-level AUROC in 2D tasks and point-level AUROC in 3D tasks. Both are computed from continuous local anomaly scores and the corresponding binary annotations. For 2D evaluation, anomaly maps are aligned with the annotation grid. Pixels with nonzero ground-truth mask values are treated as positives, while all pixels in normal images are negatives. The implementation accumulates positive and negative pixel scores in 65,536 histogram bins over $[0, 1]$ to approximate pixel-level AUROC.

For 3D evaluation, pointwise anomaly scores are matched to their corresponding annotations. Real3D-AD uses the point scores produced by Geometry3D, whereas MVTec 3D-AD uses the fused Geometry3D–PointMAE responses. Evaluation on MVTec 3D-AD is restricted to valid points whose XYZ coordinates are finite and have a norm greater than 10^{-12}. Thus, although its anomaly scores are stored on a 2D grid, the evaluation domain consists of valid 3D points. All localization metrics use continuous scores, independently of the heatmap colors used for visualization.

Region-level anomaly localization. For MVTec 3D-AD, we additionally report P-PRO to evaluate the coverage of individual anomaly regions. At threshold θ, the per-region overlap is defined as

$$\text{PRO}(\theta) = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \frac{|M_\theta \cap R|}{|R|}, \tag{A4}$$

where $\mathcal{R}$ denotes the set of annotated anomaly regions, R is an individual region, M_θ is the set of valid points predicted as anomalous at threshold θ, and $|\cdot|$ denotes set cardinality. Each region contributes equally, irrespective of its size. Anomaly regions are obtained from the 8-connected components of the ground-truth grid and restricted to valid XYZ points.

The reported P-PRO is the normalized area under the PRO–FPR curve up to a false-positive rate of $\alpha = 0.30$:

$$\text{P-PRO} = \frac{1}{\alpha} \int_0^{\alpha} g(u)\, \mathrm{d}u, \qquad \alpha = 0.30, \tag{A5}$$

where u is the false-positive rate and $g(u)$ is the corresponding per-region overlap. The false-positive rate is the fraction of valid normal points incorrectly classified as anomalous. We use 200 uniformly spaced thresholds over the evaluated score range and compute the area by trapezoidal integration, with interpolation at the integration boundary.

Anomaly understanding on MMAD. We evaluate seven tasks on the full MMAD benchmark (Jiang et al., 2025): anomaly discrimination, defect classification, defect localization, defect

description, defect analysis, object classification, and object analysis. Answers are scored by comparing the predicted option with the ground-truth answer. Anomaly discrimination uses the equally weighted mean of answer accuracy on normal and anomalous samples. Defect localization measures the accuracy of answers to spatial questions and is reported separately from pixel- and point-level localization metrics. Defect description and analysis are also evaluated using multiple-choice accuracy rather than free-form text generation quality. Questions labeled *Object Structure* and *Object Details* are combined under *Object Analysis*.

For our evaluations, the overall score is the equally weighted mean of the seven task scores:

$$\text{Average} = \frac{1}{7}\sum_{k=1}^{7} A_k, \tag{A6}$$

where k indexes the seven tasks and A_k denotes the accuracy of task k under this protocol. All metrics are percentages; higher values indicate better performance.

IAD-R1, Qwen3.5-9B, and ADClaw results in Table 3 are obtained from our own evaluations. Other results are quoted from original papers or official MMAD result files. GPT-5.5 and AnomalyClaw results are taken from Jiang et al. (2026), who evaluate a 500-image subset (2,302 questions), rather than the full benchmark. GPT-4o, Gemini-1.5-Pro, Claude-3.5-Sonnet, LLaVA-NeXT, and InternVL2 follow Jiang et al. (2025) and its official result files. We retain source-reported averages, which may differ from the arithmetic mean of the displayed scores.

B Baseline Implementation Details

For the 2D and 3D detection comparisons, we evaluate baselines using publicly available implementations with dataset-specific preprocessing and evaluation adapters. Depending on the method, we construct normal feature memories, train task-specific modules, or perform inference with released checkpoints. The following subsections describe the implementation settings used for the baseline results reported in the main text.

B.1 2D Anomaly Detection Baselines

PatchCore. PatchCore (Roth et al., 2022) matches local features to a normal memory. We use Anomalib 0.7.0 with an ImageNet-pretrained Wide-ResNet50-2, extracting features from layers 2 and 3. Inputs are resized to 256×256 without center cropping. Each category uses a 10% coreset and nine nearest neighbors for scoring. The backbone is frozen (batch size 32; seed 0).

FastFlow. FastFlow (Yu et al., 2021) models normal feature distributions using 2D normalizing flows. Our Anomalib 0.7.0 configuration uses ResNet-18, 256×256 inputs, eight flow steps, a hidden ratio of 1.0, and 3×3 convolutions. Category-specific training uses a learning rate of 10^{-3}, weight decay of 10^{-5}, and at most 500 epochs, with pixel-AUROC early stopping at a patience of three epochs. The batch size is 32 and the seed is 0.

WinCLIP. WinCLIP (Jeong et al., 2023) combines normal and anomalous text prompts with multi-scale visual features. We use a public implementation in the zero-shot setting ($k = 0$), with ViT-B-16-plus-240, LAION400M E32 weights, 240×240 inputs, and window scales of 2 and 3. The adapter produces 400×400 anomaly maps and uses the maximum response as the image-level score. The batch size is 32 and the seed is 111.

AnomalyCLIP. We use the authors' implementation of AnomalyCLIP (Zhou et al., 2024a) and the released prompt checkpoint with CLIP ViT-L/14@336px. The input resolution is 518×518, and the prompt length, text prompt depth, and text embedding length are 12, 9, and 4, respectively. Local features are extracted from layer 24, with `DPAM_layer`=20 and Gaussian smoothing $\sigma = 4$. No additional target-dataset training is performed. The batch size is 8 and the seed is 111.

AnomalyR1. We use the released Qwen2.5-VL-based 3B checkpoint of AnomalyR1 (Chao et al., 2025) without additional training. The detection adapter performs single-image binary inference and

obtains continuous anomaly scores from a softmax over the first-position Yes/No logits. Inference uses BF16, scaled dot-product attention, and a maximum image pixel budget of $1280 \times 28 \times 28$. This adapter provides image-level scores only.

B.2 3D Anomaly Detection Baselines

The 3D baselines use normal feature matching or language-guided representations. Dataset-specific adapters account for the unorganized point clouds in Real3D-AD and the organized scans in MVTec 3D-AD. Real3D-AD predictions are associated with the original points, whereas MVTec 3D-AD predictions retain the scan-grid correspondence.

BTF. We evaluate dataset-specific geometric variants of BTF (Horwitz & Hoshen, 2023). On Real3D-AD, BTF-Raw uses mean-centered XYZ coordinates, training/test sampling strides of 100/500, and all selected normal features without coreset reduction. Scoring uses Euclidean distances with three-neighbor reweighting, followed by 1-NN propagation to the full point cloud. On MVTec 3D-AD, BTF uses FPFH, voxel scale 0.05, the same sampling strides, and 224×224 maps.

PatchCore-FPFH and PatchCore-PointMAE. These variants adapt PatchCore (Roth et al., 2022) to 3D data using the Real3D-AD implementation (Liu et al., 2023), constructing normal memories from FPFH descriptors or pretrained PointMAE features. Both use a 10% coreset setting and a memory size capped at 10,000 vectors. On Real3D-AD, point clouds are centered and aligned using RANSAC. FPFH uses 33-dimensional descriptors with a voxel scale of 0.5; PointMAE uses 16,384 groups of 128 points and features from layers 3, 7, and 11. Both employ exact FAISS 1-NN scoring, with PointMAE group scores propagated to the original points by nearest-center assignment. On MVTec 3D-AD, both use five-neighbor scoring; FPFH features are pooled to a 28×28 grid, and PointMAE uses 128 points per group.

M3DM-PointBERT and M3DM-PointMAE. We use the point-cloud branch of M3DM (Wang et al., 2023) with pretrained PointBERT or PointMAE features. On Real3D-AD, both configurations use 16,384 groups of 128 points, features from layers 3, 7, and 11, and category-specific memories with 10% coreset sampling and sparse random projection parameter $\epsilon = 0.9$. Predictions are returned to the original point cloud. On MVTec 3D-AD, the configurations use 1,024 groups of 128 points, a 56×56 feature grid, and 224×224 output maps.

PointAD. PointAD (Zhou et al., 2024b) uses multi-view renderings and point–pixel correspondences with CLIP ViT-L/14@336px. Both dataset configurations use nine views, 15 training epochs, a batch size of 4, and a learning rate of 10^{-3}. The Real3D-AD adapter trains one prompt model per category and evaluates it using the corresponding checkpoint. Its feature layers are 6, 12, 18, and 24, with prompt depth 9, context length 12, text context length 4, `DPAM_layer`=20, and smoothing $\sigma = 4$. For MVTec 3D-AD, the source categories are `carrot`, `cookie`, and `dowel`.

GS-CLIP. GS-CLIP (Deng et al., 2026) combines CLIP representations with geometric and depth information. The configurations use CLIP ViT-L/14@336px, nine RGB views, nine depth views, and point–pixel correspondences. The Real3D-AD adapter uses 3,000 input points for its PointNet branch and loads the stage-one epoch-15 and stage-two epoch-10 checkpoints. Point scores are averaged over visible views, and the object score equally weights the maximum point score and the mean view score. For MVTec 3D-AD, the two training stages use 15 and 10 epochs, learning rates of 0.002 and 0.0005, respectively, and a batch size of 4.

BTP. BTP (Li et al., 2026a) uses pretrained ULIP-2 PointBERT and OpenCLIP representations. The Real3D-AD configuration normalizes each point cloud and samples 2,048 points using farthest point sampling. Features are extracted from layers 2, 5, 8, and 11. Prompt depth, context length, and text context length are 9, 8, and 4, respectively. Following the original cross-category evaluation protocol, we train each model on annotated data from a single source category and evaluate it on the remaining categories. No data from the target evaluation categories are used for training.

C IMPLEMENTATION DETAILS OF ADCLAW

C.1 KNOWLEDGE REPRESENTATION AND NORMAL MEMORY

Object ontologies and normality rules. The structured normality prior links inspection targets to their normality requirements through object, part, and attribute identifiers. Object ontologies specify constituent parts, hierarchical relations, observable attributes, and grounding prompts, while normality rules define inspection conditions, required evidence, and verification procedures. For example, the screw-bag ontology represents screws, nuts, and washers as multiple instances within the packaging, with quantity and structural rules checking the expected composition and component integrity, respectively. For the `diamond` category in Real3D-AD, the ontology defines the table, crown, girdle, pavilion, and culet, and establishes part correspondences using axial positions and radial profiles of normal templates. Table C1 presents representative examples.

Table C1: Examples of normality requirements and their inspection evidence for 2D and 3D objects.

Object	Normality requirement	Inspection evidence
Screw bag	One long screw, one short screw, two nuts, and two washers	Instance segmentation, orientation-normalized lengths, and component counts
Breakfast box	Cereal, banana chips, and almonds are jointly present and follow the normal compartment layout	Part occupancy, relative positions, and regional semantics
3D diamond	The table remains planar, the girdle is continuous, and the pavilion gradually tapers toward the culet	Registered part-level surface residuals and normal geometric profiles

Rules participate in verification only when the corresponding grounding, registration, and visibility conditions are satisfied. For 3D inspection, evaluation is restricted to observed surfaces; unobserved regions are not directly treated as evidence of missing parts.

Normal memory. All pretrained model parameters remain fixed. Normal reference samples construct visual memories and provide calibration statistics under the dataset-specific protocols in Appendix A. The 2D memory comprises global, latent local, and semantic part levels, with 16 global prototypes and 64 latent local prototypes per category, and up to 24 local prototypes per grounded part. These prototypes support reference routing, whereas dense anomaly responses are obtained by matching patch features against normal memory. The 3D PointMAE memory uses four normal templates per category on Real3D-AD and 96 normal scans per category on MVTec 3D-AD.

C.2 TOOL CONTRACTS AND VISUAL CONFIGURATION

Tool contracts specify inputs, outputs, applicability conditions, and failure handling for the principal tools in Table C2, guiding executable action selection according to current evidence requirements.

On Real3D-AD, Geometry3D establishes correspondences between the query and normal templates using FPFH, RANSAC, and point-to-plane ICP. On MVTec 3D-AD, it evaluates organized XYZ observations against normal geometric profiles. Both configurations measure deviations on observed surfaces and apply validity checks before using the geometric evidence. PointMAE uses a frozen encoder to compare local and global query features against normal memory. On MVTec 3D-AD, additional object-level scoring branches, including Uni3D, are used according to the category configuration. VirtualView renders the same scan and reports view coverage and responses.

C.3 2D AND 3D INSPECTION PROCEDURES

2D inspection. The planner forms candidate hypotheses from the input observations and normality prior, and selects inspection actions to address missing evidence. For example, screw-bag inspection first obtains a global anomaly response and part masks, then projects the cached response onto the target parts and combines composition and structural checks to revise the diagnosis. Existing feature

Table C2: Principal inspection tools and their input–output specifications.

Tool	Inputs and inspection targets	Main outputs
SAM3	Image and ontology-derived part prompts	Part masks, bounding boxes, instance counts, and confidence scores
DINOv3	Image, normal visual memory, and an optional inspection region	Distances to normal features, anomaly response maps, and associated references
CLIP	Image or local region with semantic descriptions	Semantic evidence concerning object identity, attributes, and composition
DWT	Image or local region with normal frequency statistics	Subband statistics and local texture deviations
Geometry3D	Query point cloud, normal templates, and geometric calibration settings	Registration quality, surface residuals, point-level responses, and calibrated scores
PointMAE	Query point cloud and normal point-cloud feature memory	Local and global feature distances and point-level feature responses

maps and segmentation results are reused to avoid redundant computation. The inspection budgets are 12 steps for MVTec AD and VisA, and 18 steps for MVTec LOCO AD and GoodsAD.

3D inspection. The 3D procedure begins with a geometric anomaly inspection objective and iteratively selects complementary views to gather additional evidence. The system first obtains measurements from tools such as Geometry3D. Based on tool evidence, candidate-view statistics, and summaries of previously selected views, the planner then decides whether to invoke PointMAE, inspect another virtual view, or terminate inspection. Geometric descriptions may be updated as evidence accumulates, but the procedure primarily examines existing candidate anomaly regions rather than repeatedly switching between hypotheses. Each sample uses up to six 256×256 virtual views, all reprojected from the same scan rather than newly acquired.

Evidence reuse and output. The Inspection Workbench associates each evidence item with a unique identifier, an inspection target, a spatial scope, and dependencies, while retaining masks, response maps, and reference matches for subsequent use. Hypothesis-Guided Evidence Reasoning accounts for the roles of evidence items when checking target alignment, cross-evidence consistency, and completion of required checks. After locking the numerical decisions and spatial results, the explanation module generates an evidence-grounded diagnosis.

Numerical evidence verification. For a hypothesis h, $\mathcal{E}_h$ contains the latest eligible scored record from each applicable tool. Eligible records must be successfully executed, calibrated, and assigned positive quality. The tool-supplied quality q_e is clipped to $[0, 1]$; predefined tool weights apply only to the selected decision evidence $\mathcal{D}_h$. Unscored artifacts remain available in the workbench but do not increase the numerical support count. The required support m is the larger of the planner's minimum-evidence setting and the scorer's minimum-support setting. A hypothesis with no eligible scored evidence remains insufficient rather than being accepted.

In Equation 6, σ_h is the population standard deviation of calibrated scores in $\mathcal{D}_h$ (zero for one record). The spatial agreement ι_h is the minimum IoU between the first available evidence bounding box and each remaining box, while ρ_h is the frequency of the most common semantic label. Spatial and semantic checks apply only when at least two corresponding observations are available; otherwise, their indicator factors are treated as satisfied. Equation 7 describes hypothesis acceptance, whereas image- and object-level scores follow the dataset-specific scoring branches described above.

D Additional Results

D.1 Category-level 2D results

Table D1 reports category-level results for MVTec AD, VisA, MVTec LOCO AD, and GoodsAD, covering all test images in the dataset copies used for evaluation.

MVTec AD shows strong performance across most categories, with lower image-level discrimination on screw and wood. VisA achieves I-AUROC above 90% in every category. On MVTec LOCO AD, the image-level results are more consistent than the pixel-level results, with screw bag and splicing connector remaining challenging for localization. GoodsAD exhibits substantial variation in image-level discrimination despite P-AUROC exceeding 95% across categories. These results illustrate that sample-level detection and dense localization assess distinct capabilities.

Table D1: Category-level 2D detection and localization results (%). N denotes the number of test images. I/P denote image/pixel AUROC; F1 denotes image-level F1-max.

Category	N	I	P	F1
MVTec AD				
bottle	83	99.84	99.24	99.21
cable	150	99.79	99.19	97.85
capsule	132	99.24	99.33	99.09
carpet	117	100.00	99.58	100.00
grid	78	99.83	98.79	99.13
hazelnut	110	100.00	99.18	100.00
leather	124	100.00	99.46	100.00
metal_nut	115	100.00	99.51	100.00
pill	167	98.39	98.18	97.18
screw	160	80.98	91.94	87.72
tile	117	100.00	98.23	100.00
toothbrush	42	100.00	98.85	100.00
transistor	100	99.50	97.32	98.77
wood	79	87.63	95.87	96.00
zipper	151	100.00	98.63	100.00
Macro	1,725	97.68	98.22	98.33
MVTec LOCO AD				
breakfast_box	275	80.49	92.05	79.43
juice_bottle	330	90.29	93.78	89.30
pushpins	310	84.31	92.07	85.89
screw_bag	341	83.89	72.67	83.29
splicing_conn.	312	81.50	78.09	78.21
Macro	1,568	84.10	85.73	83.22

Category	N	I	P	F1
VisA				
candle	200	97.31	99.53	92.81
capsules	160	96.31	97.95	96.32
cashew	150	91.10	99.64	89.61
chewinggum	150	98.41	99.26	97.87
fryum	150	95.34	96.77	92.47
macaroni1	200	95.38	99.73	95.21
macaroni2	200	90.48	99.38	83.10
pcb1	200	93.36	99.76	89.27
pcb2	200	94.13	98.43	90.82
pcb3	201	97.85	99.04	91.20
pcb4	201	98.32	98.81	97.89
pipe_fryum	150	95.13	99.58	94.27
Macro	2,162	95.26	98.99	92.57
GoodsAD				
cigarette_box	429	89.28	95.23	87.49
drink_bottle	781	72.78	98.50	73.83
drink_can	293	93.56	96.44	87.11
food_bottle	604	82.17	98.48	82.39
food_box	397	78.05	96.06	80.47
food_package	483	81.80	97.35	76.81
Macro	2,987	82.94	97.01	81.35

D.2 CATEGORY-LEVEL 3D RESULTS

As shown in Table D2, Real3D-AD exhibits larger category-level variation in object-level detection, while point-level localization remains comparatively stable across classes. Its results cover the full 1,206-sample test split (604 normal and 602 anomalous samples), and the Macro row reports the unweighted mean across categories. On MVTec 3D-AD, object-level performance is strong on categories such as cookie, carrot, rope, and potato, whereas cable gland, foam, and tire are more challenging. Point-level localization is generally strong, although P-PRO varies more noticeably across categories. These results indicate that object-level detection and local anomaly localization respond differently to object category and geometric structure.

Table D2: Category-level 3D anomaly detection results on Real3D-AD and MVTec 3D-AD (%). N denotes the number of evaluated samples; the Real3D-AD Macro row covers the full 1,206-sample test split. O-AUROC and O-AP measure object-level anomaly detection, while P-AUROC and P-PRO evaluate point-level anomaly localization. P-PRO is reported only for MVTec 3D-AD.

Category	N	O-AUROC	O-AP	P-AUROC	P-PRO
Real3D-AD					
airplane	100	69.26	69.89	81.64	–
candybar	100	83.20	85.57	83.35	–
car	100	78.80	81.48	81.74	–
chicken	106	64.17	70.61	78.04	–
diamond	100	98.50	98.42	84.20	–
duck	100	82.45	77.12	80.85	–
fish	100	95.20	96.11	84.81	–
gemstone	100	70.87	65.89	82.53	–
seahorse	100	72.78	78.02	86.76	–
shell	100	95.75	90.27	87.06	–
starfish	100	63.18	70.18	87.02	–
toffees	100	84.88	90.38	85.03	–
Macro	1,206	79.92	81.16	83.59	–
MVTec 3D-AD					
bagel	110	86.83	96.67	97.72	93.54
cable_gland	108	70.77	88.91	87.24	61.22
carrot	159	97.01	99.38	97.79	95.20
cookie	131	99.13	99.77	92.71	92.16
dowel	130	94.38	98.64	79.08	67.54
foam	100	64.31	87.14	90.08	66.50
peach	132	81.35	95.01	97.57	92.75
potato	114	93.38	98.48	98.20	94.96
rope	101	95.56	98.27	94.69	84.84
tire	112	59.22	79.59	93.15	76.22
Macro	1,197	84.19	94.19	92.82	82.49

E Qualitative Results

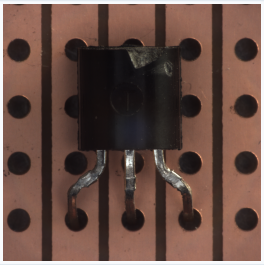
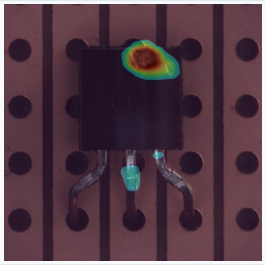
We present ten cases with input observations, tool-derived heatmaps, annotations, and evidence-checked diagnostic summaries. The summaries are edited for clarity; predictions are unchanged. Ground truth is shown only for comparison. Calibrated deviation scores measure departure from normal references, not diagnostic correctness probabilities.

2D inspection. Figures E1–E2 contrast localized defects, a missing-label violation, and a normal control. The bottle's composition rule provides evidence beyond its global appearance score, whereas the food package remains normal despite visible DINOv3 responses. Heatmaps are interpreted jointly with calibration and verification outcomes.

3D inspection. Figures E3–E4 illustrate geometric and feature evidence. Real3D-AD displays Geometry3D point scores; MVTec 3D-AD displays fused Geometry3D–PointMAE responses. These renderings visualize the evaluated scans, not additional physical observations. Object-level scores and point-level maps may use different scoring branches.

Challenging cases. Figure E5 shows two false negatives: a VisA defect fails score and spatial checks, while the MVTec 3D-AD object-level branch misses a local defect. Separate evaluation notes distinguish annotation-based assessment from model diagnosis.

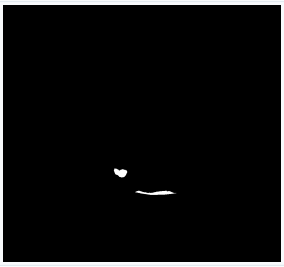
(a) MVTec AD Case 1 — Structural deviation

Input Observation	Heatmap Evidence	Ground Truth
		
RGB image	DINOv3 response	Binary anomaly mask

ADClaw | Diagnostic Results — Anomaly detected

The transistor is classified as anomalous, with the strongest visual response on the package body and weaker responses around the leads. The DINOv3 calibrated deviation score is approximately 1.0000, exceeding the normal threshold of 0.8785. The geometry hypothesis passes the score, evidence-completeness, and spatial-consistency checks. Three connected response regions are retained, linking the accepted structural diagnosis to the inspected component and its localized visual evidence.

(b) VisA Case 1 — Surface texture

Input Observation	Heatmap Evidence	Ground Truth
		
RGB image	DINOv3 response	Binary anomaly mask

ADClaw | Diagnostic Results — Anomaly detected

A texture anomaly is identified on the cashew kernel. Surface irregularities and a distinct brown spot are highlighted by the DINOv3 response relative to normal references. The calibrated deviation score is approximately 1.0000, above the threshold of 0.9482, and the hypothesis passes the evidence-completeness and spatial-consistency checks. Five connected response regions are retained. Together, the calibrated visual response and the target-aligned localization support a surface-texture diagnosis for the inspected kernel.

Figure E1: **2D structural and surface anomalies:** (a) a damaged MVTec AD transistor and (b) an anomalous VisA cashew, with DINOv3 heatmaps, annotations, and evidence-checked diagnoses.

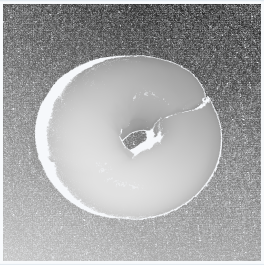
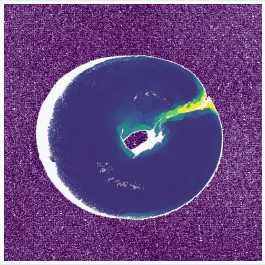
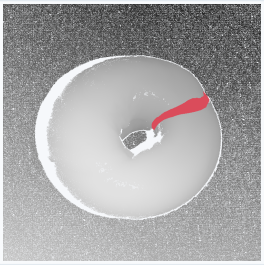


Figure E2: **Logical inspection and a normal control:** (a) a juice bottle with missing labels from MVTec LOCO AD and (b) a normal food package from GoodsAD. The all-zero annotation panel indicates the absence of an annotated defect in the normal control.



Figure E3: **Real3D-AD examples:** (a) car and (b) fish. Geometry3D heatmaps visualize deviations from normal templates; red points denote annotated anomalies in the ground truth.

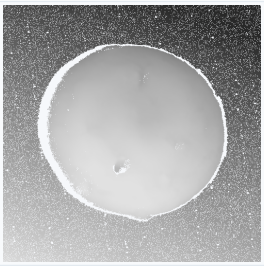
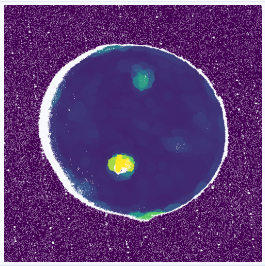
(a) MVTec 3D-AD Case 1 — Local surface irregularity

Input Observation	Heatmap Evidence	Ground Truth
		
XYZ depth rendering	Geometry3D + PointMAE	Annotated region (red)

ADClaw | Diagnostic Results — Anomaly detected

The bagel is classified as anomalous. PointMAE yields a calibrated feature-deviation score of 0.9934, exceeding its support threshold of 0.9503, whereas the Geometry3D score is 0.2689. The fused point-level heatmap highlights a localized surface region. The final object-level score is 0.9821, above the decision threshold of 0.5000. This case illustrates that feature and geometric responses can differ in strength while contributing distinct inspection evidence.

(b) MVTec 3D-AD Case 2 — Localized depression

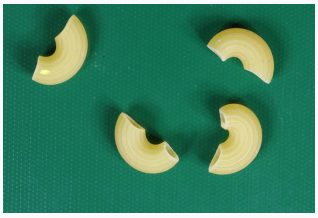
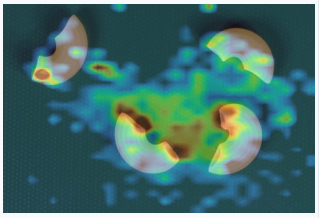
Input Observation	Heatmap Evidence	Ground Truth
		
XYZ depth rendering	Geometry3D + PointMAE	Annotated region (red)

ADClaw | Diagnostic Results — Anomaly detected

A localized depression is identified on the cookie surface. The Geometry3D validity check passes. The sink tail mean is 0.00116, exceeding the normal-profile threshold of 0.00071 and providing the key geometric evidence. PointMAE also yields a high calibrated feature-deviation score of 0.9988, above its support threshold of 0.9456. The heatmap shows the spatial response of the fused localization branch. The final object-level score is 0.9037, above the decision threshold of 0.5000.

Figure E4: **MVTec 3D-AD examples:** (a) bagel and (b) cookie. Fused point-level responses accompany XYZ renderings and annotations; diagnoses distinguish geometric and feature evidence.

(a) VisA Case 2 — Missed anomaly

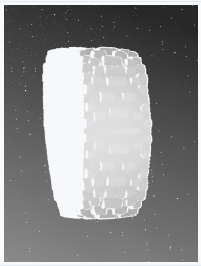
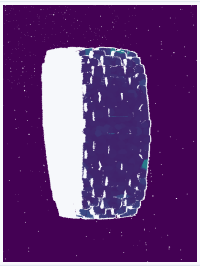
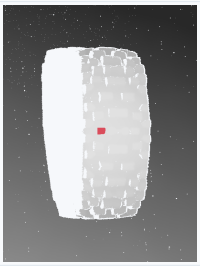
Input Observation	Heatmap Evidence	Ground Truth
		
RGB image	DINOv3 response	Binary anomaly mask

ADClaw | Diagnostic Results — No anomaly detected

The system classifies the macaroni image as normal. The DINOv3 calibrated score of 0.8359 is below the threshold of 0.8716, and the geometry hypothesis fails the spatial-consistency check. The auxiliary rule response does not change this decision. No reliable defect region is accepted despite the scattered heatmap responses.

Evaluation note. Ground truth contains a localized defect. The below-threshold score and failed spatial check result in a missed anomaly.

(b) MVTec 3D-AD Case 3 — Missed anomaly

Input Observation	Heatmap Evidence	Ground Truth
		
XYZ depth rendering	Geometry3D + PointMAE	Annotated region (red)

ADClaw | Diagnostic Results — No anomaly detected

The system classifies the tire as normal. Its Uni3D-based object-level score is 0.0103, below the threshold of 0.5000. The Geometry3D validity check passes, but the bulge and sink statistics remain within their normal-profile limits. PointMAE also gives a low calibrated score of 0.175. These measurements retain the normal object-level decision.

Evaluation note. The annotation marks a local defect. The point-level response does not lead to an anomalous object-level decision.

Figure E5: **Missed anomalies:** (a) macaroni from VisA and (b) tire from MVTec 3D-AD. Both predictions remain normal; the annotations and separate evaluation notes identify the missed defects.