HIVE-AD: ADAPTIVE MULTILEVEL FEATURE FUSION FOR TRAINING-FREE FEW-SHOT INDUSTRIAL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot industrial anomaly detection requires reliable characterization of normal patterns from only a handful of defect-free reference images, making repeated optimization for new products impractical for rapid deployment. Existing training-free methods often combine features from different layers using fixed rules and rely on a shared anomaly response for image-level detection and pixel-level localization. These strategies can miss complementary texture and structural cues, magnify harmless variations in normal samples, and produce fragmented anomaly maps. To overcome these limitations, we introduce HIVE-AD, a training-free framework that adaptively fuses multilevel features from a frozen DINOv2 backbone. HIVE-AD constructs an independent normal memory for each selected layer, determines each layer's contribution from spatial variation in normal support images, and adopts separate scoring procedures for image-level ranking and pixel-level localization. The visual pipeline requires no parameter updates, prompt tuning, or additional training for individual product categories. An optional frozen MLLM further translates localized anomaly regions into structured defect reports. In the one-shot setting, HIVE-AD achieves image-level and pixel-level AUROC scores of 97.8% and 97.6% on MVTec-AD, and 94.5% and 98.2% on VisA, respectively. Compared with a frozen DINOv2-G baseline, HIVE-AD improves image-level AUROC by 2.3 and 2.8 percentage points on MVTec-AD and VisA, demonstrating that adaptive multilevel feature fusion and task-specific scoring improve few-shot anomaly detection without additional training.

1 INTRODUCTION

Visual anomaly detection is a central problem in industrial inspection, where small defects such as scratches, contaminations, missing material, or local deformations can cause production failures (Chandola et al., 2009; Pang et al., 2021). In practice, normal samples are limited, defect appearances are long-tailed, and newly introduced product categories often require rapid adaptation. MVTec-AD and VisA reflect this setting by evaluating image-level recognition and pixel-level localization across diverse object and texture categories (Bergmann et al., 2019; Zou et al., 2022). Few-shot industrial anomaly detection (IAD) is therefore important for practical deployment, but the detector must infer normality from only a few defect-free reference images. A practical solution should adapt to a new category without repeatedly fine-tuning a category-specific model. The central challenge is to determine which frozen visual evidence is reliable and how it should be used for different prediction outputs. Instead of treating a frozen encoder as a single feature extractor, HIVE-AD adaptively combines multilevel frozen features into task-specific anomaly evidence.

Existing IAD methods model normality through reconstruction, feature distributions, memory banks, or vision-language prompts (Schlegl et al., 2019; Defard et al., 2021; Cohen & Hoshen, 2020; Jeong et al., 2023; Li et al., 2024). Training-based approaches typically optimize a category-specific reconstruction model, feature distribution, memory representation, or prompt, then treat deviations from learned normality as anomalies. Although effective, this learn-then-compare route often requires repeated training, learned memories, prompt adaptation, or auxiliary data. Such requirements hinder rapid deployment to a new category from a small normal support set.

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.

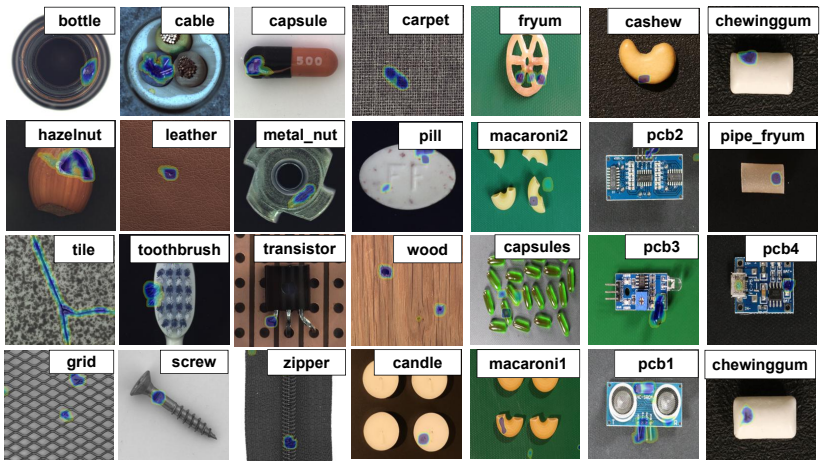


Figure 1: Qualitative localization results of HIVE-AD on MVTec-AD and VisA. Blue overlays denote high-response anomaly regions predicted by HIVE-AD rather than ground-truth annotations. Examples clearly demonstrate its ability to localize diverse texture and structural defects from only a few normal references.

Training-free methods remove optimization but retain comparison. They freeze a pretrained encoder, construct normal references from support patch memories, statistics, or subspaces, and score query deviations. DINOv2 provides dense transferable patch representations, which recent methods exploit through patch matching or subspace modeling (Oquab et al., 2023; Damm et al., 2025; Lendering et al., 2026). Without learned adaptation, however, evidence organization becomes critical. Existing pipelines commonly use fixed layer aggregation and reuse one response for image-level scoring and pixel-level localization. This is restrictive because Transformer depths encode different mixtures of appearance and structure, while ranking needs reliable abnormal peaks and localization needs spatially continuous confidence. Under few-shot support, these fixed choices can amplify benign normal variation, suppress defect cues, or fragment anomaly maps. The shift from learned models to frozen comparison motivates a hierarchical view of visual evidence. Once optimization is removed, the central problem becomes how to organize evidence. Representation depth determines which deviations are reliable, decision granularity determines how they should be read out for image-level and pixel-level predictions, and spatial extent determines how local responses form a coherent map. This formulation shifts the focus from learning another category-specific model to extracting more value from multi-level frozen evidence while accommodating different defect scales and object structures.

Based on this view, we propose HIVE-AD, a training-free framework for adaptive multilevel feature fusion that operates on frozen DINOv2 features. Given a query image and a few normal support images, HIVE-AD constructs layer-wise normal memory banks and computes patch deviations using nearest-neighbor matching. Variance-Guided Layer Weighting (VGLW) uses spatial variation within the normal support features to adapt the layer profile. Decoupled Pixel-Ranking Scoring (DPRS) provides task-specific image-level ranking and pixel-level scoring, while Pixel-Level Soft Spatial Aggregation (PSSA) consolidates local responses into a continuous anomaly map. The visual pipeline is deterministic and requires no parameter updates, prompt tuning, or category-specific training. Localized regions can additionally be forwarded to a frozen MLLM together with the query image and selected normal reference crops. The MLLM produces a structured report containing the anomaly decision, coarse location, defect type, and an evidence-based rationale, extending few-shot IAD from quantitative prediction to inspection-oriented defect reporting.

Our contributions are summarized as follows:

- We formulate training-free few-shot IAD as the organization of frozen visual evidence across representation depth, output granularity, and spatial extent.
- We develop a deterministic pipeline that combines spatial-variation-guided VGLW, task-specific DPRS, and spatial PSSA over frozen DINOv2 features.

- Experiments on MVTec-AD and VisA show consistent improvements over the bare DINOv2-G baseline without parameter updates. A frozen MLLM further converts localized evidence into structured defect reports.

2 Related Work

2.1 Classical and Feature-Based Anomaly Detection

Industrial anomaly detection has been studied from statistical, reconstruction, and representation-learning perspectives (Chandola et al., 2009; Pang et al., 2021). Early deep methods reconstruct normal samples and use reconstruction residuals as anomaly evidence, including autoencoder-based models, GAN-based methods, and structural-similarity objectives (Schlegl et al., 2019; Akcay et al., 2018; Bergmann et al., 2018). Later approaches improve reconstruction sensitivity through synthetic anomalies, restoration objectives, or defect-aware training, such as CutPaste and DRAEM (Li et al., 2021; Zavrtanik et al., 2021). Other methods model normality in feature space using Gaussian distributions, normalizing flows, teacher-student discrepancies, or feature distillation (Defard et al., 2021; Yu et al., 2021; Deng & Li, 2022; Wang et al., 2021; You et al., 2022; Batzner et al., 2024). Zero-shot feature-statistics methods such as ACR further show that distributional cues can be extracted without category-specific training (Li et al., 2023a). In few-shot settings, SPADE and PatchCore store normal patch descriptors and detect anomalies through nearest-neighbor retrieval (Cohen & Hoshen, 2020; Roth et al., 2022). Although effective, these methods often require explicit training, large or learned feature memories, or fixed feature aggregation, which can limit their flexibility under strict few-shot constraints.

2.2 Vision-Language and Multimodal Anomaly Models

Large-scale foundation models have advanced visual and multimodal representation learning. Vision-language encoders such as CLIP and ImageBind establish transferable visual-semantic spaces, while generative models such as BLIP-2 and InternVL/InternVL2 connect visual representations with language generation (Radford et al., 2021; Girdhar et al., 2023; Li et al., 2023b; Chen et al., 2024b;a). Representative anomaly-specific methods, including WinCLIP, AnomalyCLIP, AdaCLIP, PromptAD, and AnomalyGPT, explore prompt matching, prompt learning, and multimodal reasoning for anomaly detection and explanation (Jeong et al., 2023; Zhou et al., 2024; Cao et al., 2024; Li et al., 2024; Gu et al., 2024). MLLMs can also inspect query images together with localized candidate crops and normal reference regions to produce structured descriptions of anomaly decisions, locations, defect types, and visual rationales. Such reports provide an inspection-oriented interface that complements numerical scores and localization maps. However, existing approaches often depend on prompt adaptation, auxiliary data, visual-text alignment, or language inference during anomaly decision-making.

2.3 Few-Shot and Training-Free Anomaly Detection

Few-shot anomaly detection methods differ in how they characterize normal variation. Training-free approaches typically rely on normal patch retrieval, parametric feature distributions, or prompt-based semantic matching. Recent DINOv2-based methods, including AnomalyDINO and SubspaceAD, demonstrate that frozen foundation features can provide strong anomaly evidence with only a few normal references (Damm et al., 2025; Lendering et al., 2026). These methods reflect a broader shift toward reducing supervision and avoiding category-specific optimization. Nevertheless, existing training-free pipelines commonly use fixed layer aggregation or a shared response for different prediction outputs, leaving spatial-variation-guided layer weighting and task-specific evidence readout largely unexplored. HIVE-AD follows this direction by organizing frozen evidence across representation depth, decision granularity, and spatial extent, while providing an interface for converting localized evidence into structured defect reports.

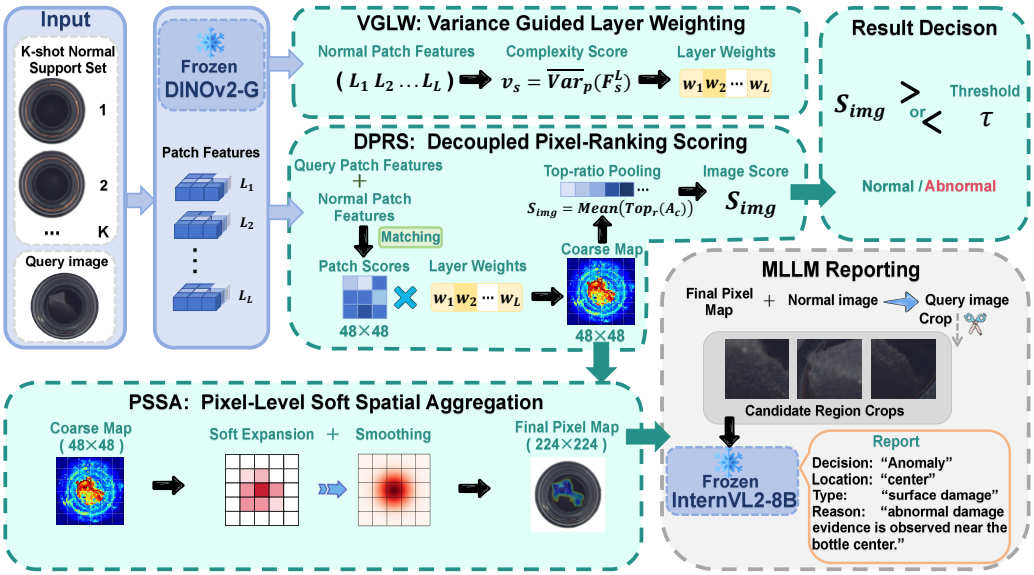


Figure 2: Overview of HIVE-AD. A frozen DINOv2-G encoder extracts multilevel patch features from normal support images and the query image. The hierarchical visual evidence core constructs layer-wise normal memories, adapts the layer aggregation through VGLW, and reads out the resulting evidence through DPRS and PSSA for image-level ranking and pixel-level localization. The localized regions may optionally be passed to a frozen MLLM for evidence-grounded reporting.

3 METHOD

3.1 PROBLEM FORMULATION

Given a query image I_q and a K-shot normal support set $\mathcal{S}_c = \{I_s^1, \dots, I_s^K\}$ from category c, HIVE-AD computes an image-level anomaly score $S_{\text{img}}(I_q)$, a pixel-level anomaly map M_q, and an optional structured report R_q. Only normal support images are used, and no model parameters are updated.

Let ϕ denote the frozen DINOv2 encoder. For each selected layer l, we denote its patch features by $F_l(I) \in \mathrm{R}^{P \times C_l}$, where P is the number of patch tokens and C_l is the feature dimension. At layer l, the small, support-derived normal memory bank is

$$\mathcal{M}^l = \{F_l(I_s^k)_p \mid I_s^k \in \mathcal{S}_c,\ p = 1, \dots, P\}. \tag{1}$$

For query patch $F_l(I_q)_p$, anomaly evidence is computed by cosine KNN matching:

$$\begin{aligned} \mathcal{N}_p^l &= \mathcal{N}_{k_{\text{nn}}}\left(F_l(I_q)_p, \mathcal{M}^l\right), \\ d_p^l &= \frac{1}{k_{\text{nn}}} \sum_{m \in \mathcal{N}_p^l} \left(1 - \cos\left(F_l(I_q)_p, m\right)\right). \end{aligned} \tag{2}$$

where $\mathcal{N}_p^l$ denotes the k_{nn} nearest normal features of query patch p at layer l. The resulting distances provide dense visual evidence for all quantitative image-level and pixel-level metrics, while the whole visual branch remains frozen and deterministic.

Overview. In this section, we present HIVE-AD as a training-free hierarchical visual evidence framework operating on frozen features, as illustrated in Fig. 2. HIVE-AD first forms representation-level evidence from layer-wise patch deviations in a frozen DINOv2 encoder (Oquab et al., 2023). VGLW adapts the aggregation of this evidence using spatial variation within the normal support features. DPRS decouples image-level ranking from pixel-level scoring, while PSSA consolidates the localization response into a continuous anomaly map. We finally describe the thresholded decision protocol and the optional frozen-MLLM reporting interface for practical deployment.

3.2 VARIANCE-GUIDED LAYER WEIGHTING

Different DINOv2 layers capture complementary cues: earlier selected blocks tend to preserve finer appearance variation, while later blocks provide more stable structural information. Rather than assuming a single layer profile for every category, HIVE-AD estimates normal-support spatial variation from the deepest selected features:

$$v_s = \frac{1}{KC_L} \sum_{k=1}^{K} \sum_{c'=1}^{C_L} \mathrm{Var}_p \left(F_L(I_s^k)_{p,c'} \right). \tag{3}$$

A larger v_s indicates greater spatial variation within the normal support feature maps and selects a later-block emphasis, while a smaller value selects an earlier-block emphasis for fine-grained evidence. Here, Var_p is taken over patch positions, so v_s measures spatial feature variation within the normal support images rather than variation across support images. We compute v_s directly from the original frozen patch features without patch-wise ℓ_2 normalization; feature normalization is applied only to the cosine KNN matching in Eq. (2). This statistic is computed only from normal support features, so the layer weighting does not require anomaly labels, category-specific calibration, or category-specific training. It therefore provides a deterministic rule for organizing evidence across the selected representation depths.

The spatial-variation score is mapped to one of three preset layer-weight templates:

$$\mathbf{w}(v_s) = \begin{cases} \mathbf{w}^{\mathrm{early}}, & v_s < \tau_1, \\ \mathbf{w}^{\mathrm{balanced}}, & \tau_1 \le v_s \le \tau_2, \\ \mathbf{w}^{\mathrm{late}}, & v_s > \tau_2. \end{cases} \tag{4}$$

This rule is deterministic and uses no learned parameters. The early and late templates form monotonic mirror profiles around uniform aggregation, providing moderate depth emphasis without allowing a single block to dominate; all templates and thresholds are fixed globally. We use five selected DINOv2 blocks with $\mathbf{w}^{\mathrm{early}} = [1.5, 1.2, 1.0, 0.5, 0.2]$, $\mathbf{w}^{\mathrm{balanced}} = [1, 1, 1, 1, 1]$, and $\mathbf{w}^{\mathrm{late}} = [0.2, 0.5, 1.0, 1.5, 2.0]$. The coarse patch-level anomaly map is computed as

$$A_c(p) = \frac{1}{\sum_{l=1}^{L} w_l} \sum_{l=1}^{L} w_l d_p^l. \tag{5}$$

For a 672×672 input and a 14-pixel patch stride, A_c has spatial size 48×48.

3.3 DECOUPLED PIXEL-RANKING SCORING

HIVE-AD uses task-level decoupled readouts because stable image scoring and sharp heatmaps require different evidence statistics. This task-level separation is the second stage of the hierarchical visual evidence pipeline. Let $U(\cdot)$ denote bilinear upsampling from the patch grid to the input resolution:

$$\tilde{A}_q = U(A_c). \tag{6}$$

For image-level ranking, HIVE-AD applies smoothing and top-ratio pooling:

$$S_{\mathrm{img}} = \frac{1}{|\Omega_r|} \sum_{u \in \Omega_r} G_{\sigma_r}(\tilde{A}_q)(u), \tag{7}$$

where Ω_r contains the top $\lceil r|\Omega| \rceil$ pixels in $G_{\sigma_r}(\tilde{A}_q)$ and Ω denotes all pixel locations. Here, G_{σ_r} denotes Gaussian smoothing in the ranking branch, which reduces isolated noisy responses before top-ratio pooling. The ratio r controls how many high-response pixels contribute to the image score, providing a balance between sensitivity to small defects and robustness to noise. This score focuses on the most abnormal regions rather than averaging over the whole image, which is important for small localized defects. The localization branch is processed separately, so the ranking score can be made robust without over-smoothing the final anomaly map.

3.4 Pixel-Level Soft Spatial Aggregation

For localization, HIVE-AD uses a separate pixel branch to consolidate spatial evidence. The upsampled anomaly map is first smoothed as

$$B_q = G_{\sigma_p}(\tilde{A}_q), \tag{8}$$

where σ_p controls localization-branch smoothing. Then, HIVE-AD combines the smoothed map with a local max-filtered response:

$$M_q = R_{224}\left((1-\alpha)B_q + \alpha\,\mathrm{MaxFilt}_h(B_q)\right). \tag{9}$$

Here, R_{224} resizes the map to 224×224, h is the max-filter window size, and α controls soft spatial expansion. The max-filter term takes the local maximum within an $h \times h$ neighborhood, allowing nearby weak responses to be reinforced by stronger anomaly evidence. PSSA connects fragmented high-response regions while preserving continuous anomaly confidence values. Unlike hard binarization or morphological expansion, this refinement keeps the map score-continuous and therefore remains compatible with pixel-level AUROC and PRO evaluation.

3.5 Thresholded Decision Protocol

For thresholded decision metrics, HIVE-AD estimates a class-specific threshold from normal support scores and synthetic perturbations:

$$\begin{aligned} a_c &= Q_\rho\left(\{S_{\mathrm{img}}(\tilde{I}_j)\}_{j=1}^{N_{\mathrm{syn}}}\right), \\ m_c &= \max_{I_s \in \mathcal{S}_c} S_{\mathrm{img}}(I_s). \end{aligned} \tag{10}$$

where $\tilde{I}_j$ is generated by randomly rotating and relocating a normal-image crop and adding small colored ellipses, and Q_ρ denotes the ρ-th percentile. These perturbations are used only for threshold estimation and introduce no parameter updates. The final threshold is

$$\tau_c = \max(a_c - \delta,\ m_c + \delta,\ \tau_{\min}), \tag{11}$$

where δ is a safety margin and $\tau_{\min}$ prevents a degenerate low threshold. The binary prediction is

$$\hat{y}_q = \mathbf{1}[S_{\mathrm{img}}(I_q) \geq \tau_c]. \tag{12}$$

AUROC and AUPR are computed from S_{img}, while pixel-level metrics are computed from M_q; therefore this threshold affects only the optional binary decision. Zero-shot and oracle threshold diagnostics are reported transparently in the supplementary material and are not used in the main comparison.

3.6 Optional Semantic Evidence Projection

The reporting branch is activated only after the visual branch localizes candidate regions. HIVE-AD selects high-response regions in M_q, extracts candidate crops from the query image, and provides them with the marked query image and a normal reference crop to a frozen MLLM, following the general visual-to-language design of modern multimodal systems (Li et al., 2023b; Liu et al., 2023; Chen et al., 2024a). The MLLM produces a structured report with a coarse location, a coarse visual category, and a short evidence-based rationale. This optional semantic projection expresses localized visual evidence in a compact, inspection-oriented form; the visual scores and maps are obtained before report generation.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate HIVE-AD on MVTec-AD and VisA, two industrial anomaly detection benchmarks (Bergmann et al., 2019; Zou et al., 2022). MVTec-AD contains 15 object and texture categories, while VisA contains 12 categories with more complex backgrounds and diverse anomaly patterns. We report image-level AUROC and AUPR, pixel-level AUROC, and PRO.

Setup	Method	MVTec-AD				VisA			
		I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
0-shot	WinCLIP (Jeong et al., 2023)	91.8	96.5	85.1	64.6	78.1	81.2	79.6	56.8
1-shot	SPADE (Cohen & Hoshen, 2020)	81.0	90.6	91.2	83.9	79.5	82.0	95.6	84.1
	PatchCore (Roth et al., 2022)	83.4	92.2	92.0	79.7	79.9	82.8	95.4	80.5
	WinCLIP (Jeong et al., 2023)	93.1	96.5	95.2	87.1	83.8	85.1	96.4	85.1
	PromptAD (Li et al., 2024)	94.6	97.1	95.9	87.9	86.9	88.4	96.7	85.1
	AnomalyGPT (Gu et al., 2024)	94.2	–	95.4	–	90.0	–	96.4	–
	KAG-prompt (Tao et al., 2025)	95.8	–	96.2	–	91.6	–	97.0	–
	IIPAD (Lv et al., 2025)	94.2	97.2	96.4	89.8	85.4	87.5	96.9	87.3
	AnomalyDINO (Damm et al., 2025)	96.6	98.2	96.8	92.7	87.4	89.0	97.8	92.5
	SubspaceAD (Lendering et al., 2026)	97.1	98.6	97.5	94.5	93.2	93.0	98.2	**95.5**
	HIVE-AD (Ours)	**97.8**	**99.0**	**97.6**	**94.6**	**94.5**	**95.3**	**98.2**	93.4
2-shot	SPADE (Cohen & Hoshen, 2020)	82.9	91.7	92.0	85.7	80.7	82.3	96.2	85.7
	PatchCore (Roth et al., 2022)	86.3	93.8	93.3	82.3	81.6	84.8	96.1	82.6
	WinCLIP (Jeong et al., 2023)	94.4	97.0	96.0	88.4	84.6	85.8	96.8	86.2
	PromptAD (Li et al., 2024)	95.7	97.9	96.2	88.5	88.3	90.0	97.1	85.8
	AnomalyGPT (Gu et al., 2024)	95.5	–	95.6	–	90.3	–	96.8	–
	KAG-prompt (Tao et al., 2025)	96.6	–	96.5	–	92.7	–	97.4	–
	IIPAD (Lv et al., 2025)	95.7	97.9	96.7	90.3	86.7	88.6	97.2	87.9
	AnomalyDINO (Damm et al., 2025)	96.9	98.2	97.0	93.1	89.7	90.7	98.0	93.4
	SubspaceAD (Lendering et al., 2026)	97.5	98.8	97.8	94.9	93.8	93.4	98.3	**95.7**
	HIVE-AD (Ours)	**98.3**	**99.3**	**97.9**	**95.3**	**95.4**	**95.6**	**98.5**	94.4
4-shot	SPADE (Cohen & Hoshen, 2020)	84.8	92.5	92.7	87.0	81.7	83.4	96.6	87.3
	PatchCore (Roth et al., 2022)	88.8	94.5	94.3	84.3	85.3	87.5	96.8	84.9
	WinCLIP (Jeong et al., 2023)	95.2	97.3	96.2	89.0	87.3	88.8	97.2	87.6
	PromptAD (Li et al., 2024)	96.6	98.5	96.5	90.5	89.1	90.8	97.4	86.2
	AnomalyGPT (Gu et al., 2024)	96.3	–	96.2	–	90.1	–	97.1	–
	KAG-prompt (Tao et al., 2025)	97.1	–	96.7	–	93.3	–	97.7	–
	IIPAD (Lv et al., 2025)	96.1	98.1	97.0	91.2	88.3	89.6	97.4	88.3
	AnomalyDINO (Damm et al., 2025)	97.7	98.7	97.2	93.4	92.6	92.9	98.2	94.1
	SubspaceAD (Lendering et al., 2026)	98.0	99.0	97.9	95.1	94.7	94.3	98.4	**96.0**
	HIVE-AD (Ours)	**98.5**	**99.3**	**98.0**	**95.5**	**95.3**	**95.4**	**98.6**	94.7

Table 1: Comparison with representative IAD methods on MVTec-AD and VisA. All results are reported in percentages. Compared results are taken from corresponding papers, and unavailable metrics are marked with dashes. The best reported values are shown in bold.

Variant	MVTec-AD				VisA			
	I-AUROC	I-AUPR	P-AUROC	PRO	I-AUROC	I-AUPR	P-AUROC	PRO
DINOv2-G	95.5	96.3	95.1	91.8	91.7	92.4	96.6	91.6
+ VGLW	95.9	97.5	96.4	92.6	92.1	93.0	97.5	92.2
+ DPRS	96.2	97.7	96.3	92.4	93.4	93.4	97.6	92.5
+ PSSA	96.1	97.5	95.8	93.7	93.0	92.6	97.2	93.2
VGLW + DPRS	**97.8**	98.7	97.3	93.9	93.8	94.7	97.8	92.4
VGLW + PSSA	97.5	98.1	97.2	93.6	93.0	94.4	98.0	93.2
DPRS + PSSA	97.5	98.5	97.5	94.0	**94.5**	95.0	97.8	93.3
Full HIVE-AD	**97.8**	**99.0**	**97.6**	**94.6**	**94.5**	**95.3**	**98.2**	**93.4**

Table 2: Ablation of HIVE-AD components on MVTec-AD and VisA for improved anomaly detection and localization.

Implementation details. The visual branch uses frozen DINOv2-G (Oquab et al., 2023), while InternVL2-8B (Chen et al., 2024a) is enabled only for optional reporting. Images are resized to 672×672, and patch features are extracted from blocks $\{11, 17, 23, 29, 39\}$. We evaluate $K \in \{1, 2, 4\}$ using only normal supports, with $k_{\text{nn}} = 1$, $(\tau_1, \tau_2) = (0.8, 1.5)$, $(\sigma_r, r) = (4.0, 0.002)$, and $(\sigma_p, h, \alpha) = (8.0, 31, 0.6)$. These values form one global configuration shared by both datasets, every category, all shot settings, and all DINOv2-S/B/L/G variants, without parameter updates or category-specific adjustment. Optional binary-decision diagnostics use $N_{\text{syn}} = 64$ boundary probes and $(\rho, \delta, \tau_{\min}) = (5\%, 0.01, 0.05)$; their full definitions and results are deferred to the supplement.

4.2 Main Results

HIVE-AD improves over SubspaceAD by 0.7 image-level AUROC points on MVTec-AD and 1.3 points on VisA in the 1-shot setting, and exceeds AnomalyDINO by 7.1 points on VisA image-level AUROC, as shown in Table 1. Across evaluated few-shot settings, HIVE-AD achieves the strongest image-level AUROC and AUPR on both benchmarks and obtains the best or tied-best pixel-level AUROC. The main exception is VisA PRO, where SubspaceAD remains higher, suggesting broader region coverage from its subspace residual maps. The optional report branch is not used to obtain these scores.

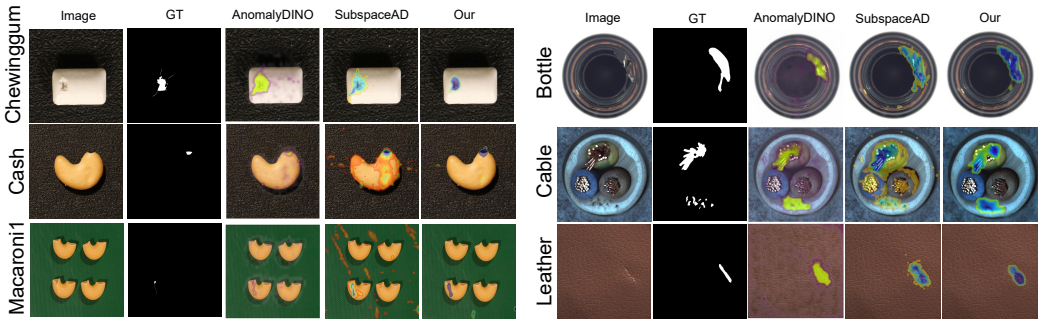


Figure 3: Qualitative comparisons on VisA and MVTec-AD. Columns show the input image, ground truth, AnomalyDINO, SubspaceAD, and HIVE-AD. HIVE-AD localizes compact high-response regions and suppresses irrelevant background activations.

Mapping	(τ_1, τ_2)	MVTec-AD			VisA		
		I-AUROC	P-AUROC	PRO	I-AUROC	P-AUROC	PRO
Discrete	$(0.6, 1.8)$	97.38	97.34	94.12	94.11	97.90	93.06
Discrete†	$(0.8, 1.5)$	97.82	97.64	94.61	94.54	98.17	93.38
Discrete	$(1.0, 1.3)$	97.82	97.64	94.61	94.27	98.05	93.23
Continuous	$(0.6, 1.8)$	97.81	97.61	94.59	94.65	98.03	93.13
Continuous	$(0.8, 1.5)$	97.83	97.62	94.55	94.29	98.06	93.24
Continuous	$(1.0, 1.3)$	97.81	97.60	94.50	94.27	98.05	93.23

Table 3: VGLW sensitivity under the 1-shot setting across both benchmarks, showing stable layer weighting. † denotes the default rule.

4.3 Ablation Study

Table 2 evaluates how HIVE-AD organizes frozen DINOv2 evidence across representation, decision, and spatial levels. The bare DINOv2-G baseline uses the same 672×672 input, blocks $\{11, 17, 23, 29, 39\}$, layer-wise normal memories, and exact cosine 1-NN matching as HIVE-AD, but uniformly aggregates layer responses, shares one coupled response for image scoring and localization, and disables VGLW, DPRS, and PSSA. The baseline achieves 95.5 image-level AUROC on MVTec-AD and 91.7 on VisA. VGLW improves depth aggregation, DPRS introduces decoupled image-level ranking and pixel-level scoring, and PSSA consolidates fragmented localization responses. DPRS alone raises VisA image-level AUROC to 93.4, while PSSA improves PRO from 91.8 to 93.7 on MVTec-AD and from 91.6 to 93.2 on VisA. Their pairwise combinations further improve multiple metrics, suggesting that the components provide complementary benefits rather than redundant post-processing. Full HIVE-AD reaches 97.8/97.6/94.6 on MVTec-AD and 94.5/98.2/93.4 on VisA for I-AUROC/P-AUROC/PRO, improving image-level AUROC over the baseline by 2.3 and 2.8 points.

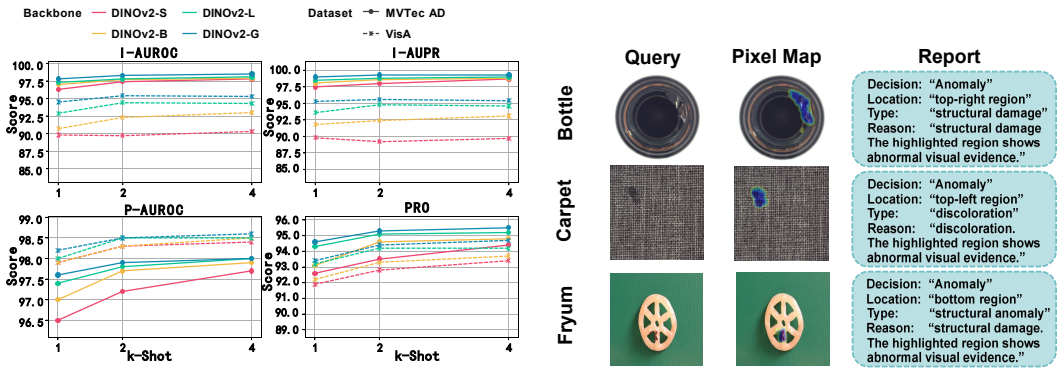
4.4 Qualitative Results

Representative qualitative comparisons on VisA and MVTec-AD are shown in Figure 3. HIVE-AD produces compact and well-aligned responses for both surface and structural defects. Compared with AnomalyDINO and SubspaceAD, it suppresses irrelevant background activations while preserving high responses around the actual defect regions. These visual results illustrate the benefit of combining depth-adaptive evidence with separate ranking and localization readouts. The resulting maps can optionally guide downstream report generation, but the visual predictions shown here are independent of the MLLM.

5 Detailed Experimental Analysis

5.1 VGLW Sensitivity

Three threshold pairs under the discrete rule and a parameter-free continuous interpolation between the same templates are compared in Table 3. Across all six settings, the I-AUROC variation is only



(a) Backbone scale **(b)** Report examples

Figure 4: Analysis of HIVE-AD backbone scaling and report generation. (a) Larger DINOv2 backbones generally improve image-level ranking while pixel-level localization remains relatively stable. (b) Qualitative examples show the query image, anomaly map, and structured MLLM report grounded in localized anomaly evidence.

0.45 points on MVTec-AD and 0.54 points on VisA. The P-AUROC variation is 0.30 and 0.27 points, while the PRO variation is 0.49 and 0.32 points, respectively. The default discrete rule achieves the best or tied-best result on four of the six dataset-metric combinations and remains within 0.11 points of the best result on the other two. Continuous interpolation provides comparable performance but no consistent improvement, so HIVE-AD retains the simpler default discrete rule.

5.2 Backbone Scale

We compare HIVE-AD with DINOv2-S, DINOv2-B, DINOv2-L, and DINOv2-G as frozen visual experts in Figure 4(a). Larger backbones generally improve image-level ranking on both datasets, indicating that richer frozen representations provide more discriminative patch evidence. DINOv2-G achieves the strongest overall performance and is therefore used as the default expert. Smaller backbones remain competitive for pixel-level AUROC, showing that the proposed evidence readout is not limited to the largest encoder and offers a practical trade-off between accuracy and efficiency.

5.3 Qualitative Report Examples

Qualitative examples from the frozen-MLLM branch are shown in Figure 4(b). Each example includes the query image, anomaly map, and a structured report grounded in localized evidence; the MLLM does not affect anomaly scores or pixel predictions. The reports provide an auxiliary human-readable summary of the visual evidence and can support rapid inspection triage. Their separation from the quantitative branch ensures that occasional language errors do not alter the detector outputs.

Additional quantitative evaluation of the generated reports is provided in the supplementary material.

6 Conclusion

We introduced HIVE-AD, a training-free few-shot IAD framework that hierarchically organizes anomaly evidence from frozen DINOv2 features without any model optimization. VGLW adapts aggregation across representation depths, DPRS decouples image-level ranking from pixel-level scoring, and PSSA consolidates spatial responses without parameter updates. On MVTec-AD and VisA, the complete visual pipeline consistently improves the bare DINOv2-G baseline and delivers strong few-shot detection and localization performance. An optional frozen MLLM converts localized evidence into preliminary structured defect reports for industrial inspection. Overall, organizing frozen visual evidence improves few-shot IAD without parameter updates while adding a transparent reporting interface.

Ethics Statement

This study evaluates HIVE-AD on the publicly released MVTec-AD and VisA industrial-image benchmarks and does not involve human subjects.

AI use statement

Generative AI tools were used for language editing and LaTeX formatting assistance. All AI-assisted changes were reviewed by the authors, who take responsibility for the final content of the manuscript.

Reproducibility statement

The method, evaluation protocol, datasets, metrics, frozen backbone configuration, selected feature blocks, and global inference hyperparameters are specified in the Method and Experiments sections. Additional implementation details, category-level results, threshold diagnostics, runtime and memory measurements, and reporting analyses are provided in the supplementary material to facilitate independent reproduction of the reported results.

References

Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pp. 622–637. Springer, 2018.

Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 128–138, 2024.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

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*, pp. 9592–9600, 2019.

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*, pp. 55–72. Springer, 2024.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. In *arXiv preprint arXiv:2005.02357*, 2020.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1319–1329. IEEE, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pp. 475–489. Springer, 2021.

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*, pp. 9737–9746, 2022.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

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*, volume 38, pp. 1932–1940, 2024.

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*, pp. 19606–19616, 2023.

Camile Lendering, Erkut Akdag, and Egor Bondarau. Subspacead: Training-free few-shot anomaly detection via subspace modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28557–28566, 2026.

Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. *Advances in Neural Information Processing Systems*, 36:40963–40993, 2023a.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9664–9674, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16838–16848, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-all few-shot anomaly detection via instance-induced prompt learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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*, pp. 14318–14328, 2022.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

Fenfang Tao, Guo-Sen Xie, Fang Zhao, and Xiangbo Shu. Kernel-aware graph prompt learning for few-shot anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7347–7355, 2025.

Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

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. *arXiv preprint arXiv:2111.07677*, 2021.

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*, pp. 8330–8339, 2021.

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*, volume 2024, pp. 49705–49737, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

SUPPLEMENTARY MATERIAL

A OVERVIEW

This supplementary material follows the protocol, terminology, and numerical results of the main paper. It provides additional implementation details, dataset-level and category-level results under the 1-shot, 2-shot, and 4-shot settings, VGLW mapping and threshold sensitivity, a decision-threshold diagnostic, an automatic audit of the optional frozen-MLLM reporting branch, additional localization examples, and additional structured defect reports. We report image-level AUROC (I-AUROC), image-level AUPR (I-AUPR), pixel-level AUROC (P-AUROC), and PRO as the principal metrics. Binary-decision accuracy is reported separately only as a threshold diagnostic. All values are percentages. All quantitative detection and localization results are produced by the frozen visual branch; the optional MLLM is used only after visual inference for report generation.

B ADDITIONAL METHOD DETAILS

B.1 EVALUATION PROTOCOL

For each category, HIVE-AD uses only normal support images under the 1-shot, 2-shot, and 4-shot settings. The query set is unchanged across the three settings. No model parameter is updated during evaluation, and no abnormal support image or test annotation is used to construct the support memory. Ground-truth masks are accessed only after inference for pixel-level evaluation and visualization. The reported implementation uses one global configuration for MVTec-AD and VisA and for every category and shot count. In particular, there is no category-conditioned override of the number of nearest neighbors, DPRS parameters, PSSA parameters, VGLW thresholds, or decision-calibration parameters. The visual score, binary decision, and pixel map are finalized before optional report generation, and no MLLM output is fed back into the quantitative pipeline.

B.2 FROZEN FEATURE EXTRACTION AND NORMAL MEMORY

HIVE-AD uses DINOv2-G as the frozen visual expert. Each image is resized to 672×672 before feature extraction, producing a 48×48 patch grid with the ViT/14 encoder. We extract features from blocks $[11, 17, 23, 29, 39]$ (zero-based indices). Multi-layer patch features are collected from the frozen encoder and ℓ_2-normalized before query-support matching. For each selected layer, the features from all normal support images are concatenated without coreset subsampling to form a category-specific normal memory. The memory is computed once for each support set and reused for every query image in the same category.

For a query patch q and a normal-memory patch m, the distance is $d(q, m) = 1 - q^\top m$ after normalization. The implementation forms the complete query–memory similarity matrix and performs exact cosine KNN matching with $k_{\mathrm{nn}} = 1$; no approximate index is used. This is the same matching rule as Eq. (2) of the main paper. The resulting layer-wise patch distances provide the dense visual evidence used by both the image-ranking and pixel-localization readouts. No feature, prompt, adapter, or threshold parameter is optimized through gradient descent.

B.3 RUNTIME AND RESOURCE ANALYSIS

We evaluate the computational cost of HIVE-AD on a single NVIDIA A40 GPU with a batch size of one. Runtime excludes model loading and is averaged over all test queries from MVTec-AD and VisA. The visual-only HIVE-AD pipeline requires approximately 0.70 seconds per image for anomaly scoring and localization. When the optional frozen MLLM is enabled, the end-to-end runtime increases to approximately 6.07 seconds per image because each prediction additionally involves autoregressive report generation.

The peak GPU memory consumption is 5,270.0 MB for the visual-only pipeline and 23,180.4 MB when the frozen MLLM is enabled. The MLLM reporting branch does not affect the anomaly scores or localization maps. It can therefore be disabled when lower latency and memory consumption are preferred, while preserving the quantitative detection and localization results.

Configuration	Time/image (s)	Peak memory (MB)
HIVE-AD visual branch	0.70	5,270.0
HIVE-AD + frozen MLLM	6.07	23,180.4

Table A1: Runtime and peak GPU memory measured on a single NVIDIA A40 GPU with batch size one. Per-image latency excludes model loading and support-memory construction.

Support-memory construction is performed once for each category and selected support set. The resulting layer-wise normal memories are cached and reused for all subsequent test queries. This one-time preprocessing operation is excluded from the reported per-image inference latency.

B.4 VGLW, DPRS, AND PSSA

VGLW selects a fixed layer-weight pattern using only the normal support set. Specifically, it computes the spatial variance of the deepest-layer patch features for each support image and averages it over support images and channels. With $\tau_1 = 0.8$ and $\tau_2 = 1.5$, VGLW uses the shallow-to-deep templates $[1.5, 1.2, 1.0, 0.5, 0.2]$, $[1, 1, 1, 1, 1]$, and $[0.2, 0.5, 1.0, 1.5, 2.0]$ for variance below τ_1, between the two thresholds, and above τ_2, respectively. The weights are normalized by their sum. Thus, layer selection depends on support statistics rather than a category name or a query image, and the selected weights remain fixed for all queries from that support set.

Decoupled Pixel-Ranking Scoring (DPRS) separates image-level ranking from pixel-level localization. The ranking branch applies Gaussian smoothing with $\sigma_r = 4.0$ followed by top-ratio pooling with $r = 0.002$, while the localization branch retains dense spatial responses for pixel evaluation. This separation avoids forcing one aggregation rule to serve both image ranking and dense localization.

Pixel-Level Soft Spatial Aggregation (PSSA) first smooths the dense anomaly map with $\sigma_p = 8.0$ and then mixes it with a local maximum-filtered map using a window size $h = 31$ and soft aggregation ratio $\alpha = 0.6$. These same values are used on both MVTec-AD and VisA, for all categories and all shot counts. PSSA remains score-valued and does not use a binary mask during refinement.

Component	Parameter	Global value
Feature extraction	input / blocks	672^2 / $[11, 17, 23, 29, 39]$
Exact matching	distance / k_{nn}	cosine / 1
VGLW	τ_1, τ_2	$0.8, 1.5$
DPRS	σ_r, r	$4.0, 0.002$
PSSA	σ_p, h, α	$8.0, 31, 0.6$
Calibration	N_{syn}, ρ	$64, 5\%$
Calibration	$\delta, \tau_{\min}$	$0.01, 0.05$
MLLM	quantitative feedback	disabled

Table A2: Complete configuration used for both datasets, every category, and all shot counts.

B.5 DECISION CALIBRATION AND FROZEN MLLM REPORTING

The thresholded decision protocol follows Eqs. (10)–(12) of the main paper. It uses $N_{\mathrm{syn}} = 64$ CutPaste-style synthetic perturbations and $\rho = 5\%$ to estimate a support-conditioned boundary. For each perturbation, a normal support image is sampled, a 10–40 pixel rectangular crop is copied to a random destination after a random rotation in $[0, 360]$ degrees, and three randomly located ellipses of size 5–15 pixels are added using black, white, red, or gray. The perturbation is generated before the common 672×672 preprocessing. These samples are simple boundary probes used only for threshold estimation; they never update the visual encoder or any other model component.

For category c, let $\widetilde{I}_j$ denote the j-th synthetic perturbation and let $\mathcal{S}_c$ be the selected normal support set. We compute

$$a_c = Q_\rho\Big(\{S_{\mathrm{img}}(\widetilde{I}_j)\}_{j=1}^{N_{\mathrm{syn}}}\Big), \qquad m_c = \max_{I_s \in \mathcal{S}_c} S_{\mathrm{img}}(I_s), \tag{13}$$

and set

$$\tau_c = \max(a_c - \delta,\ m_c + \delta,\ \tau_{\min})\,, \tag{14}$$

where $\delta = 0.01$ and $\tau_{\min} = 0.05$. The binary prediction is

$$\widehat{y}_q = \mathbf{1}[S_{\text{img}}(I_q) \geq \tau_c]\,. \tag{15}$$

AUROC and AUPR are computed directly from S_{img}, while pixel-level metrics are computed from M_q; therefore, τ_c affects only the optional binary decision. No real anomalous image or test label is used for the zero-shot calibration.

The optional reporting branch uses frozen InternVL2-8B without fine-tuning. It receives the query image, localized high-response regions, and a normal reference crop. The output contains a textual restatement of the visual anomaly decision, a coarse location, a coarse visual defect category, and a short evidence-based rationale. The MLLM is invoked only after the frozen visual branch has produced its score, decision, and map. Its output is neither fused with nor allowed to veto the visual result; therefore, enabling or disabling report generation leaves all quantitative detection and localization results unchanged.

B.6 Decision-Threshold Diagnostic

We distinguish zero-shot (ZS) accuracy from oracle accuracy. ZS accuracy uses only the normal supports and synthetic boundary probes in Eqs. (1)–(3), without test labels. Oracle accuracy selects the threshold after observing test labels and is included only as a diagnostic upper bound on score separability; it is not a deployment metric and is not used for comparison with AnomalyGPT or any other method.

Dataset	Setting	ZS Acc.	Oracle Acc.
MVTec-AD	1-shot	82.9	95.7
VisA	1-shot	73.2	88.7

Table A3: One-shot decision-threshold diagnostic. Oracle accuracy uses test labels only to characterize an upper bound on continuous-score separability and is excluded from all main claims.

B.7 Comparison Scope and Reproducibility

The comparison table in the main paper evaluates complete published systems. Unless a baseline explicitly reports the same setting, its original backbone, input resolution, and post-processing are retained; therefore, the table should not be interpreted as a fully compute-matched backbone replacement study. The DINOv2-S/B/L/G analysis in the main paper separately isolates the effect of visual-backbone scale within HIVE-AD. All HIVE-AD results in the main paper and this supplement use the single configuration in Table A2; category identities are used only to select the corresponding normal support memory, not to alter scoring or refinement rules.

C Dataset-Level Results

Table A4 reproduces the principal HIVE-AD results reported in the main paper under the unified configuration in Table A2. Ranking metrics are computed from the continuous image score and pixel metrics from the final continuous anomaly map. The optional MLLM is not used to obtain any value in this table.

Dataset	Setting	I-AUROC	I-AUPR	P-AUROC	PRO
MVTec-AD	1-shot	97.8	99.0	97.6	94.6
	2-shot	98.3	99.3	97.9	95.3
	4-shot	98.5	99.3	98.0	95.5
VisA	1-shot	94.5	95.3	98.2	93.4
	2-shot	95.4	95.6	98.5	94.4
	4-shot	95.3	95.4	98.6	94.7

Table A4: Dataset-level HIVE-AD results under the 1-shot, 2-shot, and 4-shot settings.

D CATEGORY-LEVEL RESULTS

D.1 MVTEC-AD

Table A5 reports category-level I-AUROC and P-AUROC for all 15 MVTec-AD categories. Category-level I-AUPR and PRO are not included because the supplied evaluation logs report these two metrics only at the dataset level. HIVE-AD maintains strong pixel-level localization across the three support sizes. The screw category is more difficult for image-level ranking, while its P-AUROC remains above 99 in all settings.

Category	1-shot		2-shot		4-shot	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC
bottle	100.0	98.9	100.0	98.9	100.0	98.9
cable	95.8	97.0	97.0	97.6	96.8	97.7
capsule	94.1	98.0	95.0	98.1	96.6	98.3
carpet	100.0	99.2	100.0	99.3	100.0	99.3
grid	100.0	99.2	100.0	99.3	100.0	99.3
hazelnut	97.7	99.3	98.9	99.4	99.5	99.4
leather	100.0	99.2	100.0	99.2	100.0	99.2
metal_nut	99.9	96.4	100.0	96.5	100.0	97.2
pill	95.6	94.8	96.6	95.0	96.1	95.2
screw	88.1	99.3	90.6	99.4	92.0	99.4
tile	100.0	97.3	100.0	97.3	100.0	97.4
toothbrush	98.6	98.8	98.9	98.9	98.1	98.8
transistor	98.5	93.8	99.3	94.8	99.6	95.2
wood	99.2	96.8	99.4	96.8	99.4	96.8
zipper	99.8	97.7	99.8	97.9	99.8	97.9
Mean	97.8	97.6	98.3	97.9	98.5	98.0

Table A5: Category-level results on MVTec-AD.

D.2 VISA

Table A6 reports category-level I-AUROC and P-AUROC for all 12 VisA categories. Pixel-level performance remains high across categories with complex backgrounds. Additional normal references improve several difficult categories, including fryum, pcb2, pcb4, and pipe_fryum.

Category	1-shot		2-shot		4-shot	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC
candle	94.6	99.4	94.0	99.4	94.0	99.5
capsules	96.3	98.1	95.7	98.7	96.6	98.8
cashew	97.8	98.3	96.8	98.4	97.0	98.5
chewinggum	99.3	99.6	99.3	99.5	99.6	99.6
fryum	91.0	92.7	97.5	93.9	96.5	94.4
macaroni1	90.8	98.8	93.1	99.2	92.3	99.1
macaroni2	84.8	98.1	85.4	98.7	81.6	99.0
pcb1	94.6	99.2	95.5	99.2	95.8	99.2
pcb2	90.9	98.4	91.8	98.7	93.7	98.9
pcb3	97.8	98.7	97.6	98.8	97.3	98.9
pcb4	98.6	97.8	99.1	98.0	99.5	98.3
pipe_fryum	98.0	98.9	99.0	99.0	99.4	99.1
Mean	94.5	98.2	95.4	98.5	95.3	98.6

Table A6: Category-level results on VisA.

D.3 Runtime Analysis

We report total evaluation time on a single NVIDIA A40 GPU, as shown in Table A7. The visual HIVE-AD branch, which contains the complete detection and localization pipeline, takes 269.54 minutes, lower than SubspaceAD under the same setting. This efficiency comes from caching frozen support features and direct query-support matching, without support augmentation or category-specific model fitting. Enabling the optional report interface increases the total time to 391.42 minutes because of additional visual-language inference. Thus, the main detector can be deployed independently of the reporting component.

Method	Structured report	Time (min)
SubspaceAD	No	386.22
HIVE-AD (No MLLM)	No	**269.54**
HIVE-AD	Yes	391.42

Table A7: Total evaluation time on a single NVIDIA A40 GPU.

E Automatic Report Quality Audit

A preliminary automatic audit is conducted on ground-truth anomalous samples for which the optional MLLM report is generated after a positive visual decision. Location accuracy measures exact agreement between the reported coarse location and the cell containing the ground-truth-mask centroid on a 3×3 grid. Localization–text alignment averages agreement with the anomaly-map peak cell and agreement with the mask-centroid cell. Coarse visual type is evaluated only on MVTec-AD, where dataset defect subtypes can be mapped to a manually checked seven-class visual taxonomy: structural damage, surface damage, contamination-like appearance, discoloration, deformation, foreign object, and uncertain visual anomaly. VisA does not provide a compatible defect-subtype taxonomy, so type accuracy is not reported for that dataset. The inherited anomaly-decision field is excluded from every report metric.

Dataset	Setting	Location	Text align.	Type
MVTec-AD	1-shot	62.9	60.7	65.3
	2-shot	64.5	63.2	67.5
	4-shot	65.1	62.8	67.9
VisA	1–4-shot range	57.4–62.5	60.8–65.3	–

Table A8: Automatic audit of the optional reporting branch. Type is unavailable on VisA because no compatible subtype taxonomy is provided.

Table A8 indicates moderate spatial and coarse semantic grounding. The audit does not evaluate open-ended language quality, and standardized reference reports are unavailable; therefore, the correctness and specificity of generated rationales are not independently established. These values are diagnostic rather than evidence of fully faithful explanations.

F Additional Localization Results

Figures A1–A5 show the five supplied MVTec-AD visualization pages, and Figures A6–A9 show the four supplied VisA visualization pages. Each source page presents query images, ground-truth masks, and anomaly maps under the 1-shot, 2-shot, and 4-shot settings. The examples cover surface defects, structural defects, small anomalies, repeated objects, and visually complex backgrounds.

The predicted regions remain concentrated around the annotated defects across the three support sizes. Additional support images generally preserve the main defect region while reducing isolated responses. These examples complement the aggregate results by showing the behavior of HIVE-AD across a broad range of categories. The masks are displayed only after inference for visual comparison and are never provided to the detector or the MLLM.

G Additional Structured Defect Reports

Figure A10 shows the first page of the supplied report PDF, which contains MVTec-AD examples, while Fig. A11 shows the second page with VisA examples. Each example includes the query image, ground-truth mask, localized anomaly response, and generated structured report.

The report restates the frozen visual decision and summarizes a coarse location, coarse visual defect category, and short rationale based on the highlighted evidence. These examples illustrate only the optional reporting branch. All quantitative detection and localization results are produced before report generation by the frozen visual branch. The masks shown in the figures are evaluation annotations and are not provided to the MLLM.

H Discussion and Limitations

The complete results show that HIVE-AD remains effective across different numbers of normal references. Additional support images generally improve image-level ranking on MVTec-AD and pixel-level localization on VisA. The category-level results also show that strong pixel localization can be maintained even when image-level ranking is more difficult.

HIVE-AD is slightly below SubspaceAD on VisA PRO despite its strong pixel AUROC. A plausible failure mode is under-coverage of large or fragmented anomalous regions: the compact continuous responses favored by HIVE-AD can rank anomalous pixels well while covering a smaller fraction of an irregular connected component at some PRO thresholds. Increasing α or h can expand coverage, but choosing category-specific values would reintroduce the tuning concern; we therefore retain the global $\alpha = 0.6$ and $h = 31$ setting. Since Decoupled Pixel-Ranking Scoring (DPRS) separates image scoring from localization, future multi-scale refinement can be studied without changing the image-ranking rule.

VGLW has a boundary case when support variance lies near τ_1 or τ_2, because a small change in the support set can switch discrete templates. As reported in Table 3 of the main paper, alternative thresh-

olds and continuous interpolation vary by at most 0.54 I-AUROC points, 0.30 P-AUROC points, and 0.49 PRO points across the two datasets. This reduces, but does not eliminate, the heuristic nature of the variance-to-depth mapping. Future work may estimate smoother support-conditioned weights while preserving the training-free protocol.

The reporting branch produces coarse visual descriptions rather than dataset-specific defect names. This avoids unsupported fine-grained claims but limits semantic detail. Type accuracy is reported only on MVTec-AD because VisA lacks a compatible subtype taxonomy, and the automatic audit does not establish the correctness of open-ended rationales. More precise spatial alignment and blinded human evaluation remain important future work.

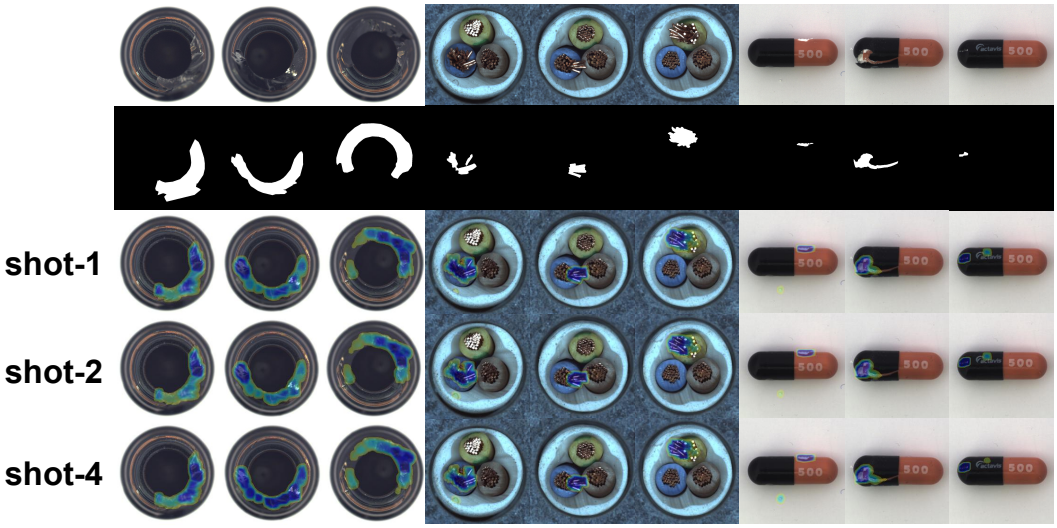


Figure A1: Additional MVTec-AD localization results, part 1. The rows show query images, ground-truth masks, and HIVE-AD anomaly maps under the 1-shot, 2-shot, and 4-shot settings.

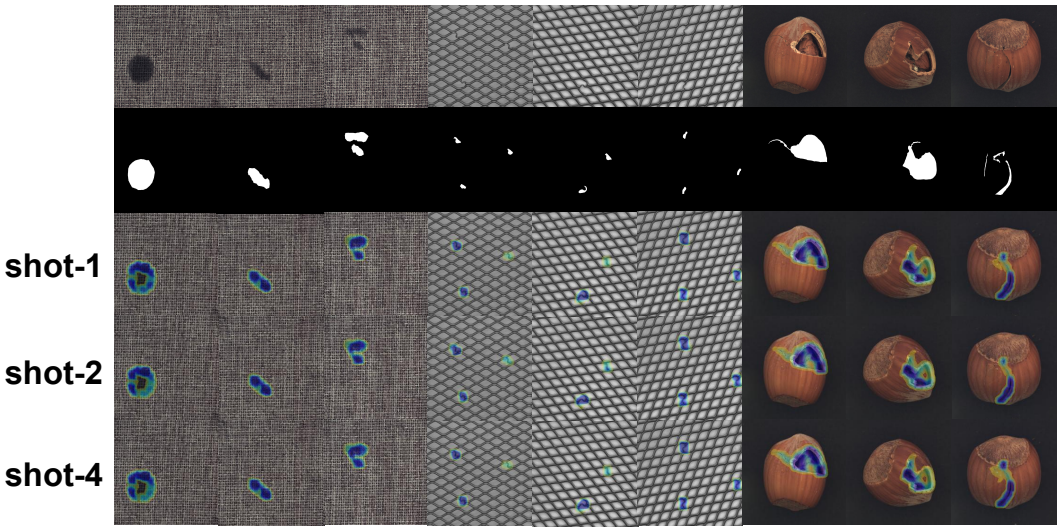


Figure A2: Additional MVTec-AD localization results, part 2.

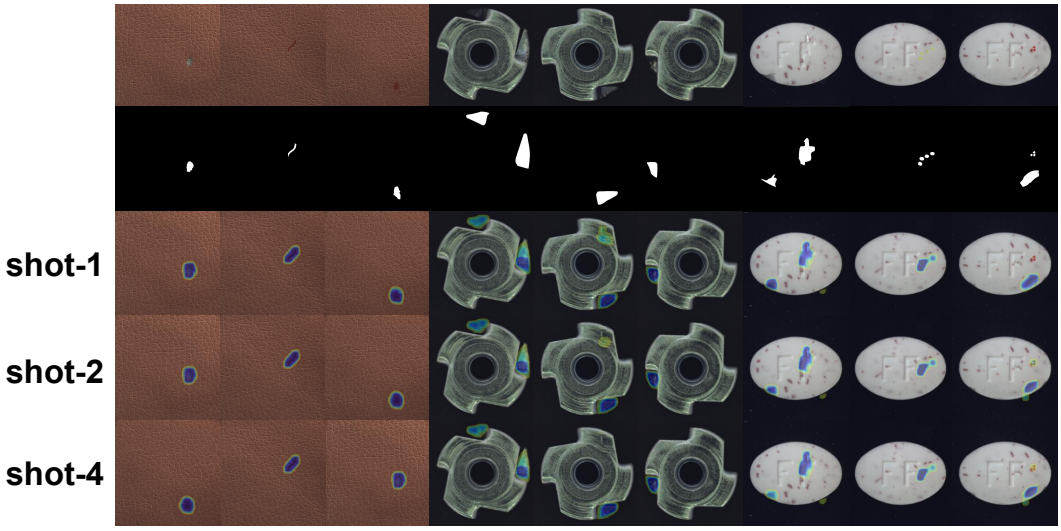


Figure A3: Additional MVTec-AD localization results, part 3.

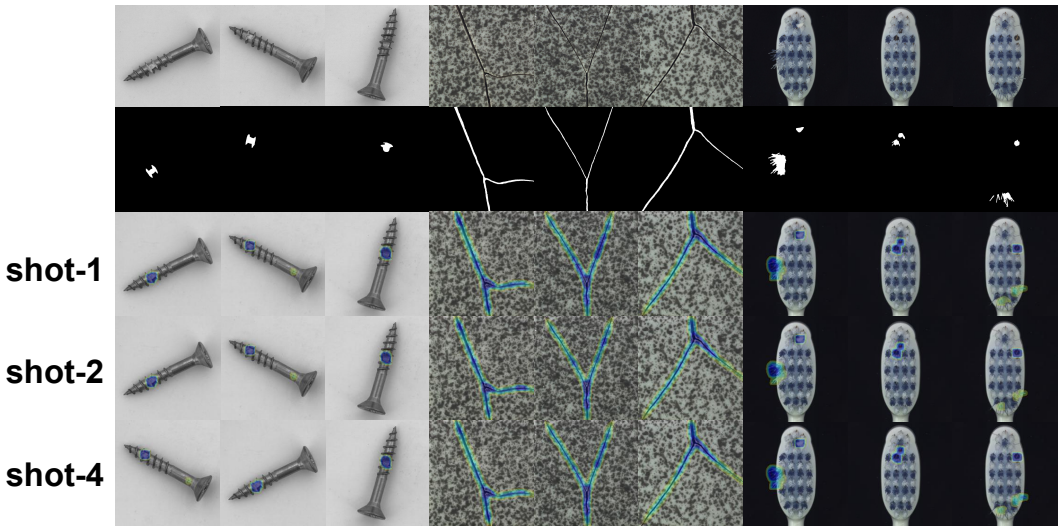


Figure A4: Additional MVTec-AD localization results, part 4.

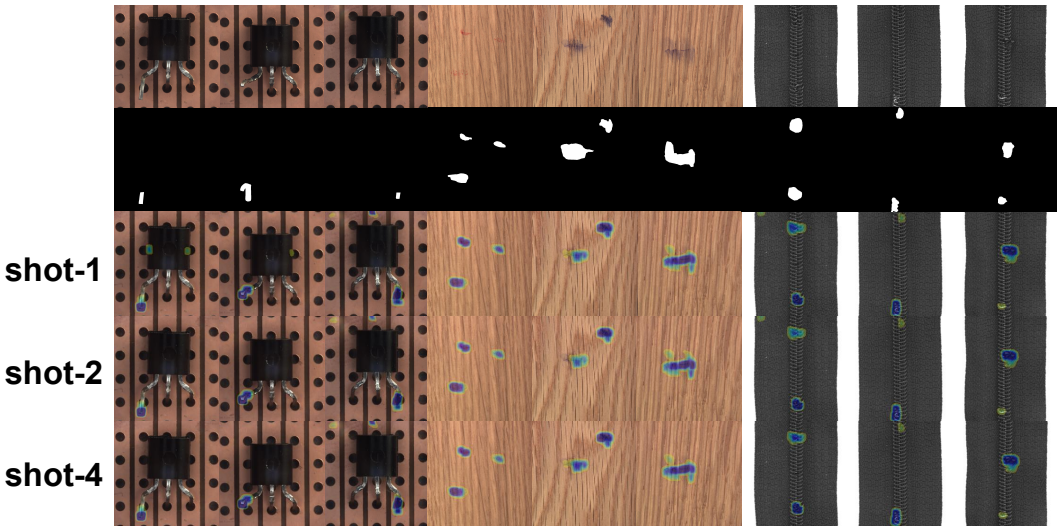


Figure A5: Additional MVTec-AD localization results, part 5.

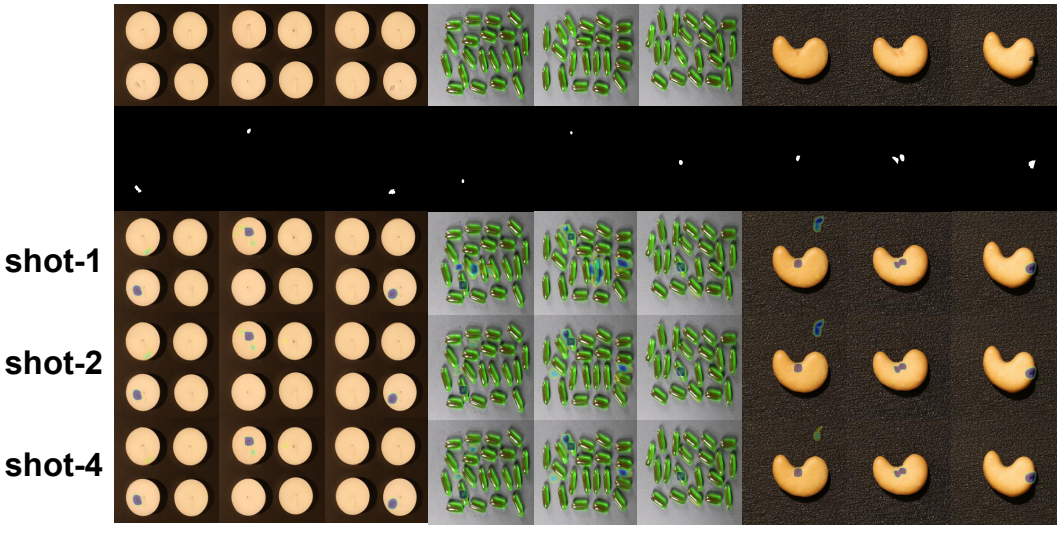


Figure A6: Additional VisA localization results, part 1. The rows show query images, ground-truth masks, and HIVE-AD anomaly maps under the 1-shot, 2-shot, and 4-shot settings.

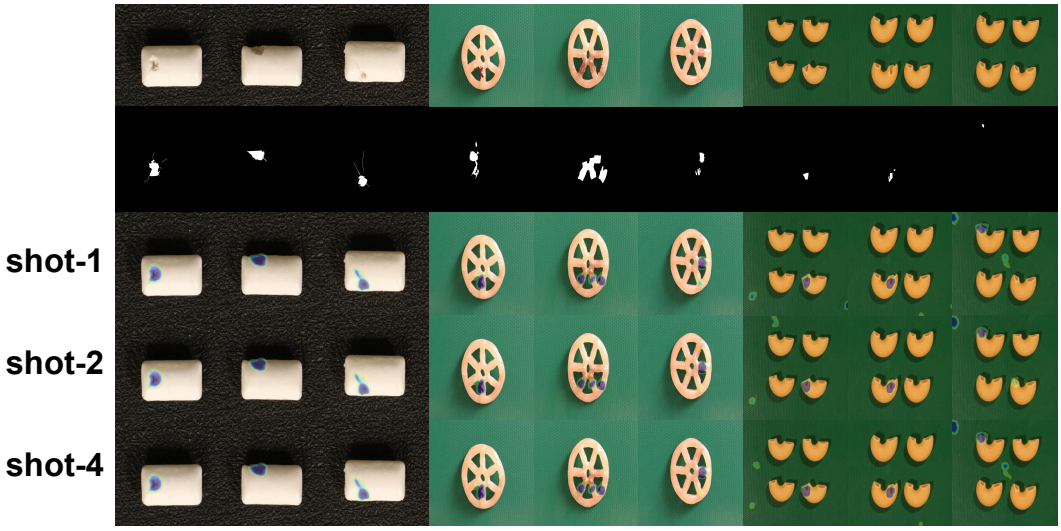


Figure A7: Additional VisA localization results, part 2.

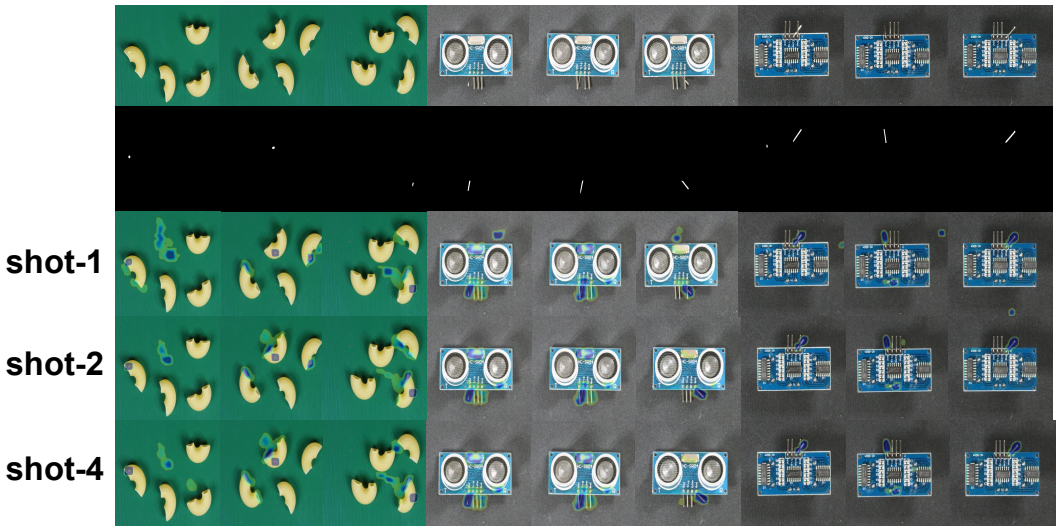


Figure A8: Additional VisA localization results, part 3.

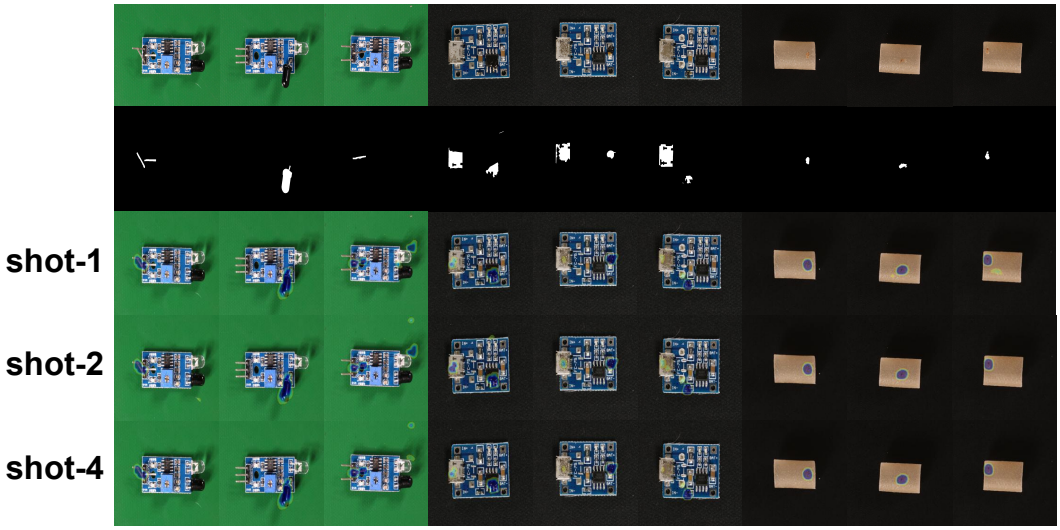


Figure A9: Additional VisA localization results, part 4.

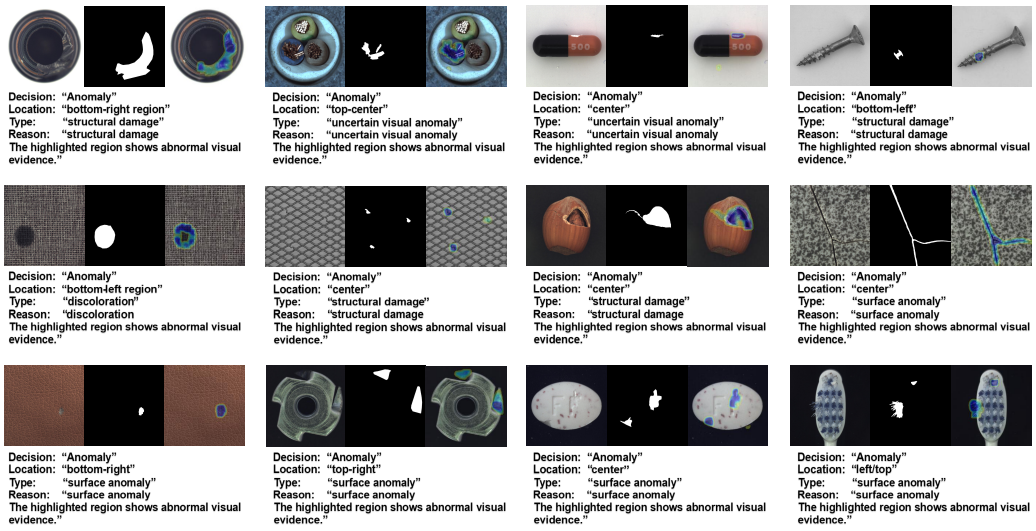


Figure A10: Additional structured defect reports on MVTec-AD. Each block contains the query image, ground-truth mask, anomaly map, and generated report.

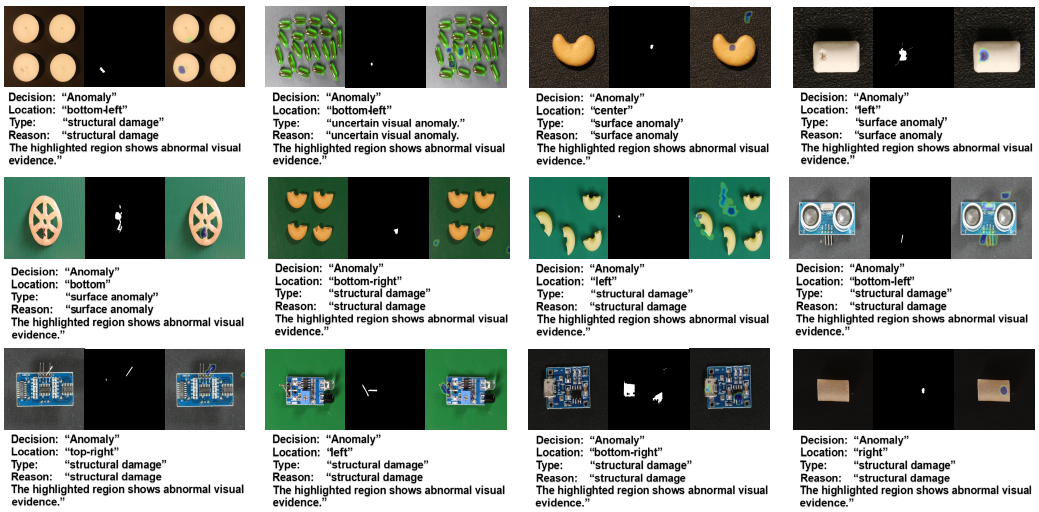


Figure A11: Additional structured defect reports on VisA. Each block contains the query image, ground-truth mask, anomaly map, and generated report.