Pay for the Reasoning Once: Compiling Agent-Written Programs for Visual Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

Agentic anomaly detectors invoke a vision-language model on every image they inspect, using tools that a human author must write in advance. We ask whether normal examples can instead support the construction and verification of reusable inspection code. ANP (Anchored Normality Programs) lets an agent compose visual primitives into executable checks, tests the candidates on held-out normal examples and controlled violations, and freezes the accepted checks with their fusion rule and calibration. On MVTec LOCO, the complementary ensemble ANP-E2 achieves 96.00% mean image-level AUROC, 8.50 percentage points above AnomalyMoE's reported result. The efficient ANP configuration achieves 94.34% mean AUROC. On the breakfast category, it requires 191.13 ms per image under our shared timing protocol with precomputed masks, corresponding to $8.78\times$ and $47.32\times$ speedups over our LogSAD and UniVAD executions. Neither configuration invokes a language model during inference. Controlled ablations of the efficient configuration attribute a gain of 2.87 AUROC percentage points to the verified program layer (95% CI $[1.93, 3.84]$); additional component ablations examine its role in the ensemble. Experiments across six benchmarks assess the scope of normal-only compilation, while acceptance and calibration audits characterise the evidence available before labelled defects are observed.

1 Introduction

A product can be defective even when every visible component looks normal. A missing item, an incorrect pairing, or a misplaced connector violates an assembly rule without necessarily producing an unusual local texture. Logical anomaly detection therefore requires a representation of the relations that make an assembly valid, often with only normal examples available for a new product (Bergmann et al., 2022a). These relations are natural inspection programs, but constructing them for each product remains difficult.

Recent agentic detectors address this problem by asking a vision–language model to inspect each image with a predefined toolbox (Zhang et al., 2026; Miao et al., 2025; Tan et al., 2026; Jiang et al., 2026). The results are real, and so is the bill. Reasoning is repeated for each sample, with reported costs of 2.9 s per image (Miao et al., 2025) or more than three model calls (Jiang et al., 2026). Without a reusable executable check, the same inspection logic is reasoned through and paid for again on subsequent images. Moreover, the inspection operations are specified in advance: adapting them to a new assembly still depends on the available human-written interface.

Our insight is that the reusable outcome of reasoning should be the inspection check itself (Figure 1). Normal examples describe what a check must accept; controlled changes to those examples provide a way to test the distinctions it makes. Together, they support an offline process that proposes executable checks, tests their behaviour, and retains useful checks for subsequent images. The central question is therefore how to turn normal examples into a reusable inspection program and determine which candidates to deploy before defect labels become available.

ANP (Anchored Normality Programs) implements this idea through a generate–verify–freeze procedure (Figure 2). An agent composes a small library of visual primitives into executable checks. A discriminative verifier tests both invariance to valid changes and sensitivity to constructed vio-

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.

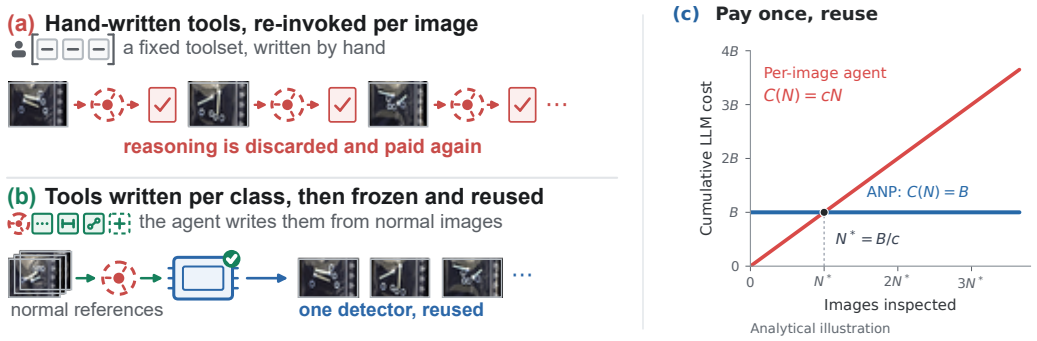

Figure 1: **Authorship of the tools and reuse of inspection reasoning.** (a) A human-written toolbox is invoked by a vision–language model for each image. (b) The agent writes class-specific checks from normal references and freezes them for reuse. (c) Cumulative LLM reasoning cost grows as cN for per-image agents and remains at the offline compilation cost B for ANP, with crossover $N^* = B/c$. This analytical illustration excludes per-image visual execution, which both approaches still require.

lations, then returns execution feedback for revision. Accepted checks are combined with visual detector scores and calibrated on a separate normal split. Once frozen, the resulting runtime needs neither an online language model nor per-image tool selection.

The complementary ensemble ANP-E2 reaches 96.00% mean image AUROC on MVTec LOCO, 8.50 points above AnomalyMoE's published result. The efficient ANP configuration reaches 94.34% mean AUROC; on breakfast it takes 104.10 ms for the model/mask-ready core and 191.13 ms from decoding to score, with 8.78× and 47.32× online speedups over our LogSAD and UniVAD executions under the shared protocol with precomputed masks. Its controlled no-program ablation gives a +2.87-point program-layer contribution, while the ensemble has its own component analysis. Across industrial, medical, 3D and video benchmarks, the framework also supports normal-calibrated detector composition. The current configurations achieve 99.88% on MVTec-AD, 99.47% on the official VisA split, and 97.45% on MVTec 3D-AD with the separately identified batch-transductive extension.

Paying for reasoning once means completing an offline construction stage for each class, which may include multiple rounds of generation and verification. The frozen program then serves subsequent images. The central verification challenge is to establish useful checks from normal examples before labelled defects are available, and to assess whether their calibration transfers to deployment data. If compilation incurs LLM cost B per class and online reasoning costs c per image, reuse amortizes that expense after $N^* = B/c$ images (Figure 1c). Here B includes all generation and verification rounds; visual execution remains a per-image cost.

The paper contributes an executable representation of class-specific inspection checks, a normal-only verifier with execution feedback, and a frozen deployment procedure with explicit calibration. Controlled ablations measure the contribution of the program layer and its individual branches, while cross-domain experiments examine the scope of the compilation framework. The accompanying acceptance and calibration analyses characterise when the normal-only evidence supports deployment.

2 Related Work

Agentic visual anomaly detection places a vision–language model within the inspection loop, typically with a plan–reason–reflect structure, a small set of image operations, and retrieval over normal references (Zhang et al., 2026; Jiang et al., 2026). These systems differ in how the loop is trained, employing prompting, supervised trajectories, or reinforcement learning with rewards that penalise excess tool calls (Miao et al., 2025; Tan et al., 2026), but they share the two properties that this paper addresses: the toolbox is written by hand before deployment, and the loop executes once per image. Fine-tuned multimodal detectors (Gu et al., 2024) avoid the tool-authorship question but inherit the

per-image model call, and efficiency work within this line reduces the number of calls rather than removing the model from the inference path.

Closer to our setting are systems that induce rules or code from a language model and subsequently execute them. LogiCode (Zhang et al., 2025b) translates expert-written logical rules into Python against a pre-implemented image API. Its rules and thresholds are supplied by a person, and its perception is read in part from human annotations, so its accuracy is not comparable with that of unsupervised baselines. AnomalyRuler (Yang et al., 2024) induces rules from normal frames but requires a language model at test time to apply them, because those rules are not executable. LogSAD (Zhang et al., 2025a) generates a rule set once and executes it with fixed detectors, without a verification loop, and LogicQA (Kwon et al., 2025) filters generated question checklists on normal images, which constitutes the closest precedent for using normals to validate generated artifacts. Our focus is the joint treatment of executable checks, normal-only verification, and calibration. Visual inspection makes this combination particularly demanding because each check must bind to a perceptual measurement rather than to a named table column.

Outside anomaly detection, agents that author their own tools are by now established. A stronger model writes reusable Python for a cheaper model to invoke (Cai et al., 2024), toolboxes grow and are pruned by execution consistency (Wang et al., 2024b), repositories are wrapped into callable tools with unit tests as the acceptance criterion (Wölflein et al., 2025), and self-evolving agents synthesise capabilities on demand (Qiu et al., 2025; Wang et al., 2024a). The recurring difficulty in that literature is determining whether a synthesised tool is correct. Normal examples provide task-specific acceptance evidence for visual inspection, and controlled transformations extend that evidence to candidate violations. This makes anomaly detection a useful setting for studying verification when positive defect examples are unavailable.

Among detectors that dispense with a language model entirely, memory-bank and reconstruction models (Roth et al., 2022; Deng & Li, 2022; Batzner et al., 2024) are fast and strong on structural defects but do not represent logical constraints, and a recent reconstruction model (Guo et al., 2025) does not report LOCO. Few-shot unified detectors (Jeong et al., 2023; Gu et al., 2025) generalise across classes at high per-image cost. The trained generalist mixture (Gu et al., 2026) attains 87.5 on LOCO without any language model and is the baseline of greatest relevance to us, because it demonstrates that unified accuracy and speed are achievable without induction. Our work adds an explicit program representation of the checks applied to a class. Additional comparisons with component-aware CSAD (Hsieh & Lai, 2024) and SALAD (Fučka et al., 2025), including source settings and aggregation details, appear in Appendix B.2. Our conformal machinery is standard (Vovk et al., 2005; Angelopoulos & Bates, 2023), and the closest use of it to gate an agent's actions is that of Ren et al. (2023) in robotics.

3 Method

We represent normality through executable predicates on visual measurements. Counts, pairings, and spatial relations complement appearance scores by expressing how individually plausible components should form a valid assembly. A *check* is a short program π over a fixed library of visual primitives (segment, count, measure, compare) that maps an image x to a scalar together with a textual reason,

$$\pi : x \mapsto (s, r), \qquad s = \pi(x) \in \mathbb{R}, \tag{1}$$

where a larger s indicates a more anomalous image. Three properties follow from this representation, and the remainder of the design rests upon them. Because π is executable, it can be tested before deployment; because it executes on normal images, that test requires no defects; and because s is a named quantity, a threshold upon it can be calibrated and a raised flag can be explained.

3.1 Label-Free Acceptance

Deployment for a new class begins with its normal images $\mathcal{N}$, which we partition into disjoint fitting, gating, and calibration sets $\mathcal{N}_{\text{fit}}, \mathcal{N}_{\text{gate}}, \mathcal{N}_{\text{cal}}$. What a check must react to, and what it must ignore, are specified together by a corruption grammar $\mathcal{G} = \mathcal{G}^+ \cup \mathcal{G}^-$ of eight operators over $\mathcal{N}_{\text{gate}}$. The four operators in $\mathcal{G}^+$ preserve validity through light jitter, recolouring, same-location cloning and small shifts; the four in $\mathcal{G}^-$ violate it by dropping, duplicating, moving or cross-recolouring components.

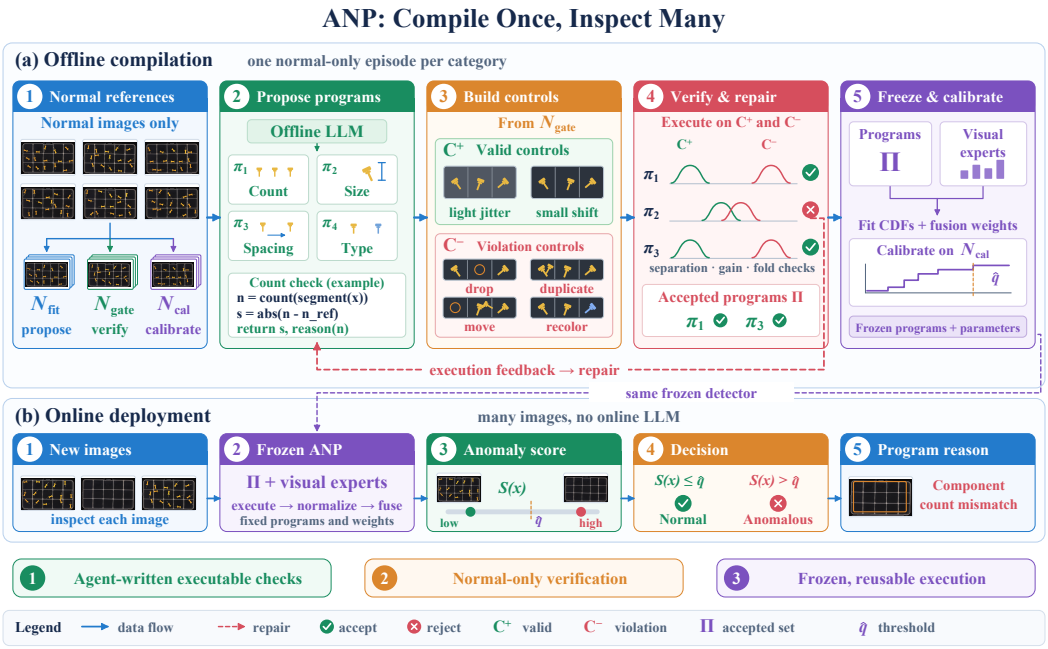

Figure 2: **Compile once, inspect many.** (a) Normal images support program proposal, control construction, verification with execution feedback, and fusion calibration on separate subsets. Candidates must separate valid from violating controls and pass gain and source-fold checks. (b) Accepted programs and visual experts form a frozen detector that returns a score and program-derived reason without an online language model. Photographs are real LOCO inputs; program sketches, controls, distributions, and outputs illustrate the mechanism. Algorithm 1 gives the procedure.

Each operator contributes 20 controls per class, yielding a valid set $\mathcal{C}^+$ and a violating set $\mathcal{C}^-$ of 80 images each, every one derived from a normal image and none from a defect.

A candidate π is discarded immediately if it raises an exception, returns a constant, or flags $\mathcal{N}_{\text{gate}}$. A survivor is scored on the two control sets,

$$a(\pi) = \text{AUROC}\Big(\{\pi(x) : x \in \mathcal{C}^+\},\ \{\pi(x) : x \in \mathcal{C}^-\}\Big), \tag{2}$$

which demands invariance to $\mathcal{G}^+$ and sensitivity to $\mathcal{G}^-$ simultaneously, so that a check firing on any change whatsoever scores no better than one firing on none. Because a score computed over 160 controls may be attained by chance, acceptance is a conjunction rather than the single inequality $a(\pi) \ge \tau$. A candidate must clear τ, must improve the class's control AUROC over the route without it, and must hold a non-negative gain at the median of five source folds and on at least three of them individually; the constants are given in Appendix A. No step consults a defect, and the entire cost is incurred once per class.

The connection between the control score and downstream detection depends on the coverage of the transformation family.

Assumption 1 (Grammar coverage)**.** *Let $\mathcal{D}$ denote the unknown distribution of real defects in the class. The acceptance score $a(\pi)$ predicts the detection score* $\text{AUROC}(\mathcal{N}, \mathcal{D})$ *only to the extent that the corruption grammar $\mathcal{G}$ resembles $\mathcal{D}$.*

The grammar therefore specifies the distinctions tested during compilation. Section 4.4 examines this relationship empirically.

3.2 Execution Feedback and Repair

A rejection does not terminate the episode, since the execution that produced it is itself evidence. The agent receives the failure kind, namely a raised exception, a constant output, a flagged normal or a

low $a(\pi)$, together with the measurements that π computed on $\mathcal{N}_{\text{gate}}$, and then proposes a revision. The loop executes to a fixed attempt budget and freezes whatever has been accepted by that point. Algorithm 1 states the episode in full.

Feedback can revise the measurement used by a check as well as its decision logic. In the semantic repair study, the agent replaces component-region measurements with two semantic queries and a typed set comparison. This branch contributes $+0.77$ mean LOCO points (Appendix B.2). Each proposal uses a short context containing the latest execution report; the appendix compares this design with accumulating the full repair history.

Algorithm 1 The ANP deployment episode for one class. No step reads a defect label.

Require: normals $\mathcal{N}$, primitive library, grammar $\mathcal{G}$, threshold τ, attempt budget T
Ensure: frozen program set Π, fusion weights w, threshold $\hat{q}$
1: partition $\mathcal{N}$ into $\mathcal{N}_{\text{fit}}, \mathcal{N}_{\text{gate}}, \mathcal{N}_{\text{cal}}$; set $\mathcal{C}^{\pm} \leftarrow \mathcal{G}^{\pm}(\mathcal{N}_{\text{gate}})$ and $\Pi \leftarrow \emptyset$
2: **for** $t = 1$ **to** T **do**
3: $\quad\pi \leftarrow \textsc{Propose}(\mathcal{N}_{\text{fit}},\ \text{library},\ \text{last report})$ ▷ short fresh-draft context
4: $\quad$**if** π raises, is constant, or flags $\mathcal{N}_{\text{gate}}$ **then** record the failure kind and measurements and **continue**
5: $\quad$compute $a(\pi)$ by (2); **if** it clears τ and the gain and source-fold conditions hold **then** $\Pi \leftarrow \Pi \cup \{\pi\}$ **else** record $a(\pi)$ as feedback
6: **end for**
7: fit $\hat{F}_j$ on $\mathcal{C}^{+}$ and search w by (3); set $\hat{q} \leftarrow S_{(k)}$ on $\mathcal{N}_{\text{cal}}$, $k = \lceil (n+1)(1-\alpha) \rceil$ ▷ Proposition 1
8: **return** $(\Pi, w, \hat{q})$, frozen

3.3 Fusion and Calibration

Accepted checks must be combined before they can be calibrated, and both the combination and its weights are fixed on the controls rather than on evaluation labels. Each constituent score s_j is placed on a common scale by its own empirical CDF $\hat{F}_j$, estimated on that class's valid controls alone, and the frozen score is a convex combination

$$S(x) = \sum_j w_j \hat{F}_j\big(s_j(x)\big), \qquad w \in \Delta, \quad w_{\text{gen}} \geq \tfrac{1}{2}, \tag{3}$$

with w searched on a $1/40$ grid and selected by $a(\cdot)$ on the same controls. The floor on w_{gen} caps the compiled checks at half the fused score, thereby bounding what a check admitted in error can cost a class; Section 4.4 is the reason we require that bound. The cross-domain extensions instead set w from the across-seed variance of each program's control score, so that a stable program outweighs an erratic one (Appendix A). Calibrating S converts an arbitrary cut-off into a stated error rate.

Proposition 1 (Calibrated false-positive rate)**.** *Let $S_1, \ldots, S_n$ be the values of the frozen score S on the n calibration images in $\mathcal{N}_{\text{cal}}$, and for a target level $\alpha \in (0,1)$ with $\alpha \geq 1/(n+1)$ set the threshold to the order statistic $\hat{q} = S_{(k)}$, where $k = \lceil (n+1)(1-\alpha) \rceil$. If a new normal image X is exchangeable with the calibration images, then $\mathbb{P}\big(S(X) > \hat{q}\big) \leq \alpha$.*

This is the standard split-conformal guarantee (Vovk et al., 2005; Angelopoulos & Bates, 2023). A deployment-side audit measures the flag rate on a second normal batch to assess whether the calibrated operating point transfers to deployment data (Section 4.5).

What is deployed is the accepted program set, the fusion rule, and one threshold per class, and the runtime loads and executes them. This freezing is the property for which the method is named, and it also distinguishes the approach from a cascade, since nothing in the inference path can fall back to a language model because none is present.

4 Experiments

4.1 Experimental Setup

We evaluate on MVTec LOCO (Bergmann et al., 2022a) for the main result, and report cross-domain results on MVTec-AD (Bergmann et al., 2019), VisA (Zou et al., 2022), BMAD (Bao et al., 2024), MVTec 3D-AD (Bergmann et al., 2022b) and Ped2 (Mahadevan et al., 2010).

The primary LOCO configuration uses normal training images and scores test images independently. FS denotes fitting on the full normal training split, TF denotes the source method's training-free setting, and 1S denotes one-shot adaptation. BT is an additional designation for methods that rank scores across the unlabelled evaluation cohort. The video and 3D extensions use this operation and are marked FS+BT. Table 2 separates published results from our executions; source-specific settings and author-checkpoint comparisons are given in Appendix B.2.

ANP denotes the complete verified-program configuration, including the expert-reference push-pins branch and agent-conditioned screw_bag repair. The no-program ablation retains its detector pipeline, features, calibration, and fusion. We also report individual branch ablations and a separate fully autonomous synthesis study. ANP-fast removes heavier primitives to measure the accuracy–cost trade-off. Accuracy and latency use the same frozen runtime binding; setup and timing boundaries are stated with the measurements. A precision variant reaches 94.48% mean AUROC; Appendix B.3 gives its implementation and repeatability study.

4.2 Performance across Benchmarks

Table 1: **Anomaly detection across six benchmarks.** Image AUROC (%) is reported, with frame AUROC for Ped2. BMAD averages BrainMRI, LiverCT, and RESC. Published rows retain their source settings; our rows use dataset-specific compiled configurations. Bold marks column maxima.

Method	Prot.	MVTec-AD	VisA	LOCO	BMAD	MVTec 3D-AD	Ped2
Unified-setting results reported in AnomalyMoE, Table 1 (Gu et al., 2026)							
UniAD	FS	87.5	80.2	71.3	75.67	72.5	72.5
ReContrast	FS	95.0	90.1	79.1	78.37	75.1	83.7
UniNet	FS	84.7	87.5	77.8	57.30	70.0	61.7
Dinomaly	FS	98.8	95.7	78.2	82.60	86.5	97.1
UniVAD	TF	97.8	93.5	71.0	78.57	80.2	94.3
LogSAD	TF	96.9	94.5	86.2	74.27	84.5	85.5
AnomalyMoE	FS	99.5	98.1	87.5	85.13	87.2	97.1
Additional published reference, source settings (Guo et al., 2025)							
Dinomaly2-L	FS^{p}	**99.9**	99.3	—	—	97.4	—
Our compiled configurations							
ANP	FS	99.88	**99.47**	94.48	**87.34**	93.27	98.57
ANP + BT	FS+BT	—	—	—	—	**97.45**	**99.36**
ANP-E2	FS	—	—	**96.00**	—	—	—

FS = full-shot; TF = training-free. ^{p}Dinomaly2-L values are published results, not our reruns. Reference budgets and backbones follow each source. The VisA entry uses the two-scale and two-area reconstruction composition (Table 20). The MVTec-AD entry uses the annealed reconstruction expert with fixed continuous three-role composition (Table 37). The base LOCO entry is ANP-P (Appendix B.3); ANP-E2 is the additional ensemble in Appendix B.4. For 3D, our FS entry uses RGB; Dinomaly2-L and our BT extension use RGB+3D. Dinomaly2-L's RGB-only result is 95.0 (its Table 4). This overview retains source settings; the entries do not constitute a matched-protocol ranking. The 3D BT entry combines the historical geometry route with the two-scale reconstruction expert (Table 44). BT uses unlabelled evaluation-cohort ranks. Paired gains and configuration details appear in Appendix B.10; LOCO-specific references appear in Tables 2 and 5.

Table 1 places the industrial, logical, medical, 3D, and video results alongside published generalist references. It retains each source's evaluation setting and our dataset-specific compiled routes. The base LOCO entry uses the precision configuration in Appendix B.3, with ANP-E2 shown as an additional ensemble. The 3D and video batch-transductive extensions are reported separately. Official-split VisA controls appear in Appendix B.10. On the official VisA split, the complete composition reaches 99.47% mean AUROC over all twelve classes. The RGB+3D extension reaches 97.45% mean AUROC with fixed two-scale reconstruction evidence, the highest 3D entry in Table 1. The complete configuration and resolution controls appear in Table 44.

The compilation framework combines compatible detector outputs under a shared normal-only selection and calibration contract. Its fixed three-role composition reaches 99.88% mean AUROC on MVTec-AD after a cosine-decay fine-tune of the reconstruction expert. Before this final fine-tune, replacing the historical reconstruction expert and calibrating continuous evidence improves the pre-

ceding configuration by 0.33 points (paired 95% CI $[+0.13, +0.64]$). The final fine-tune adds 0.01 points; its paired interval and the complete controls appear in Appendix B.11.1. Table 29 in Appendix B.10 gives the paired gains against frozen predecessors, followed by per-domain comparisons and program ablations.

4.3 Accuracy and Inference Cost

Table 2: **Image-level anomaly detection on MVTec LOCO.** Values are mean-class pooled image AUROC (%). Published results retain their source settings; reproduced results use our LOCO manifest. All ANP configurations use independent-image inference; the original configuration is timed in Table 3. The complementary ensemble ANP-E2 is defined in Appendix B.4. CSAD/SALAD author-checkpoint results, with equal logical/structural aggregation, appear in Table 5. Bold marks column maxima within this table.

Method	Prot.	Brk.	Juice	Push.	Screw	Splic.	Mean
Published results							
GCAD (Bergmann et al., 2022a)	FS	83.9	99.4	86.2	63.2	83.9	83.3
PatchCore (Roth et al., 2022)	FS	77.1	94.6	74.1	73.3	86.0	81.0
SINBAD (Cohen et al., 2023)	TF	—	—	—	—	—	86.8
AnomalyMoE (Gu et al., 2026)	FS	91.1	94.7	88.7	69.9	92.9	87.5
EfficientAD-M (Batzner et al., 2024)	FS	—	—	—	—	—	90.7
UniVAD (Gu et al., 2025)	1S	—	—	—	—	—	71.0
Our reproductions							
RD4AD	FS	67.0	94.1	72.3	66.6	81.9	76.4
EfficientAD	FS	76.9	95.7	78.2	66.2	80.1	79.4
PatchCore	FS	85.1	96.2	79.7	67.8	87.2	83.2
LogSAD (Zhang et al., 2025a)	TF	**95.7**	95.5	83.1	83.5	93.5	90.3
ANP (ours)	FS	95.32	99.36	95.43	86.97	94.64	94.34
ANP-E2 (ensemble)	FS	95.05	**99.64**	**96.78**	**91.74**	**96.81**	**96.00**

FS = full-shot; TF = training-free; 1S = one-shot (source settings). ANP-E2 combines the retained semantic/program branches with two SALAD-derived detectors trained on normals and Tiny-ImageNet auxiliary data; its score-equivalent full online breakfast execution is timed in Table 3. The ANP program-layer ablation appears in Table 4.

Table 3: **Measured inference latency.** RTX A6000, LOCO breakfast_box, batch size 1, with CUDA synchronization. Core measures model/mask-ready inference; online additionally includes decode, preprocessing, and transfer. Lower is better.

Method	LLM calls	Core mean (ms)	Online mean (ms)	n
Shared protocol with precomputed masks where applicable				
ANP	0	**104.10**	**191.13**	80
AnomalyMoE (Gu et al., 2026)	0	130.81	334.04	80
LogSAD (Zhang et al., 2025a), cached SAM	0	1360.21	1677.36	20
UniVAD (Gu et al., 2025)	0	8879.98	9044.88	10
Full RGB-to-score execution, including online mask generation				
ANP-E2	0	—	3701.77	275

The ANP row is the original configuration in Table 2; ANP-P has a separate instrumented measurement in Appendix B.3. Model/reference setup is excluded. The first block excludes pre-generated score-free masks where applicable; the separate E2 measurement generates both mask paths online and retains diagnostic map capture. All measurements cover breakfast only. E2 is paired with the full ensemble's accuracy row; its breakfast AUROC is 95.05%. Cached-SAM LogSAD is score-equivalent to direct execution (maximum $|\Delta| = 0$ at tolerance 10^{-12}).

ANP-E2 attains 96.00% mean image AUROC, the highest mean in Table 2 and 8.50 points above the published AnomalyMoE result. It combines the retained semantic and program evidence with two detector seeds using fixed weights. The efficient ANP configuration pairs 94.34% mean AUROC with 104.10 ms core compute and 191.13 ms decode-to-score latency on breakfast, using precomputed

masks (Tables 2 and 3). On breakfast, ANP achieves higher AUROC than ANP-E2 (95.32% versus 95.05%), while its measured latency is 19.37× lower (191.13 versus 3701.77 ms) with precomputed masks for ANP and online mask generation for ANP-E2. Ensemble definitions, seed comparisons and CSAD/SALAD references are given in Appendix B.4 and Table 5.

For the efficient configuration, the program layer contributes +2.87 AUROC percentage points over the otherwise identical pipeline (94.34% versus 91.48%; 95% CI [+1.93, +3.84]; Table 4). The gain is concentrated rather than uniform. On breakfast box, juice bottle and splicing connectors, the frozen configuration uses no program branch and preserves every detector score, hence the ranking (+0.00 pp each). Pushpins and screw bag improve by +10.51 and +3.83 pp, respectively (Figure 8), through checks on component counts and composition. Selective augmentation therefore improves two classes while preserving the other three, with no class-level AUROC regression in this evaluation.

Branch-level ablations and program provenance are given in Appendix B.2; Appendix B.4.1 reports the separate component analysis for ANP-E2.

Table 4: **Contribution of the verified program layer.** The efficient ANP configuration is compared with its matched no-program pipeline. Both rows use the runtime binding of Table 2, with no route or weight reselection. AUROC is averaged over the five LOCO classes; the gain and its 95% confidence interval are in percentage points (pp), using 10,000 class-stratified paired-bootstrap draws.

Configuration	AUROC (%)	Gain (pp)	95% CI (pp)
Without program layer	91.48	—	—
ANP (complete)	**94.34**	+2.87	[+1.93, +3.84]

Figure 3 illustrates how the retained programs change the ranking of two logical anomalies. Normal references make the missing components and changed composition visible. The complete classwise gains, ROC curves, and a fixed image grid appear in Appendix B.13.

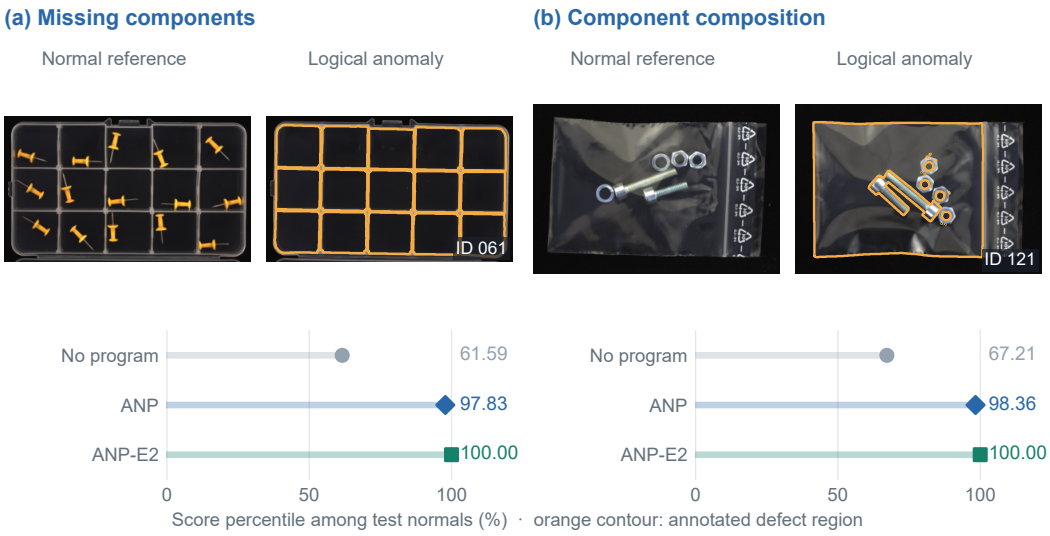

Figure 3: **Compiled checks change the ranking of logical anomalies.** Each panel pairs a normal training reference with a test anomaly and the recorded outputs of three frozen configurations. Scores are percentiles among the same class's test normals, with half credit for ties; they describe ranking, not calibrated decisions. Orange contours show dataset defect annotations. These two illustrative cases complement the aggregate ablation in Table 4; Figure 10 uses a fixed selection across all five classes.

The efficient runtime is 1.75×, 8.78× and 47.32× faster online than our AnomalyMoE, cached-SAM LogSAD and UniVAD executions, respectively, on the same GPU. Figure 4 separates core from decode-to-score timing. The separate 3.70 s ANP-E2 measurement includes both online seg-

mentation paths and diagnostic map capture; all 275 breakfast test images reproduce the ensemble's score ordering exactly.

4.4 Predictiveness of the Label-Free Acceptance Test

The verifier evaluates a check's response to constructed violations as well as its stability on normals. In the counting and relation probes, all six class–grounding combinations improve with verification, by 9.60 to 31.60 AUROC percentage points. On splicing, replacing count stability with the discriminative gate changes the accepted artifact's logical AUROC from 65.70% to $(76.40 \pm 2.40)\%$. These results support testing the distinction a check makes, beyond whether its output remains stable on normal images.

Figure 5 and the all-class ablation in Appendix B.7 also identify the scope of this evidence. In the screw_bag candidate replay, discriminative AUC ranges from 25.33% to 80.56%, while downstream all-anomaly AUROC remains between 44.73% and 58.09% (Appendix B.8). The two plotted screw_bag false accepts are expert-written and agent-written candidates: both pass the proxy task but show weak downstream discrimination. Their shared failure points to a mismatch between component-level corruptions and the length or thread differences required by real defects, rather than a failure specific to automatic code generation. This motivates improving the corruption grammar and measurement sensitivity together. The complete trajectories in Figure 6 retain the underlying execution outcomes.

4.5 Deployment-Side Calibration Audit

The deployment audit uses a second batch of normal images to check the transferred operating point. A flag rate above 2α triggers recalibration or marks the branch as untrusted. At $\alpha = 0.05$, the LOCO audit retains the original calibration for three classes and triggers for breakfast_box and pushpins. Breakfast can be recalibrated, whereas the pushpins structural branch remains untrusted. Figure 7 and Appendix B.9 give the source-specific rates and the separate fallback analysis. This audit provides a normal-only deployment diagnostic alongside the ranking metric.

4.6 Fully Autonomous Synthesis

In the autonomous study, the agent proposes checks and the gate selects them without manual intervention. Across 27 class–seed runs, 16 produce accepted programs (59.3%; Wilson CI 40.7–75.5%). This measures end-to-end synthesis acceptance over the supplied primitive library. The complete ledger, including execution failures and gate rejections, appears in Appendix B.8; downstream utility is assessed separately from acceptance.

5 Limitations

The current agent composes a supplied primitive library, and verification depends on how well normal-derived controls represent relevant violations. Extending the measurement library and the coverage of attribute-level controls is a natural next step. Conformal guarantees require exchangeability; the deployment audit assesses the operating point under new normal data. Published and author-checkpoint comparisons retain their stated settings, while the component ablations use the same frozen pipeline.

6 Conclusion

ANP turns class-level inspection reasoning into reusable executable checks. Normal examples support program generation, discriminative verification, and calibration before the runtime is frozen. The complete verified-program system reaches 94.34% mean LOCO AUROC and gains 2.87 percentage points over its matched no-program counterpart. ANP-E2 raises mean AUROC to 96.00%; the efficient configuration takes 104.10 ms for the model/mask-ready core and 191.13 ms from decoding to score on breakfast with precomputed masks. The branch ablations and cross-domain results establish concrete uses of the representation. Compiling and testing inspection logic before deployment offers a practical route to auditable anomaly detection without an online language model.

AI USE STATEMENT

A generative model is a component of the method itself rather than only an authoring aid. The deployment episode described in Section 3 prompts a vision–language model to propose executable inspection code, and every program our system accepts was written by that model or, for the expert-written references used as controls, by the authors. Section 4.4 reports what the model produced without editing, including the candidates that failed to compile and the accepted program that collapses downstream, and no generated program was repaired by hand before evaluation.

We also used a large language model while preparing the manuscript, for editing and for tightening prose in text the authors had drafted. All numbers, claims, tables and figures were produced by our own experiment code and checked against the run records by the authors; no result, citation or reference in this paper was generated by a language model. We reviewed all AI-assisted text and take responsibility for the final content of this work.

ETHICS STATEMENT

The study uses public anomaly-detection benchmarks and collects no new human subject data. The intended application is visual inspection. Executable checks make individual measurements available for inspection, and the normal-only audit assesses whether a calibrated operating point transfers to deployment data. The medical experiments evaluate benchmark transfer and do not establish clinical suitability.

REPRODUCIBILITY STATEMENT

Section 4.1 states the evaluation protocol, and Appendix B.1 records the splits, the control-set construction, and the metrics. The corruption grammar, the acceptance conjunction, and the fusion rule are specified in Section 3 with their constants given in Appendix A. Every threshold, weight and search parameter is fixed on normal-derived controls before the evaluation manifest is opened, and Appendix B.2 records the runtime binding used for both the accuracy and the latency measurements so that the two are not taken from different configurations. For the autonomous loop, Appendix B.8 gives the complete ledger of all 27 class–seed runs with their execution outcomes and scores. Section 3 also states the assumption on which the acceptance test depends (Assumption 1) and the conditions under which the calibration guarantee holds (Proposition 1).

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. BMAD: Benchmarks for medical anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. doi: 10.1109/cvprw63382.2024.00408. URL https://arxiv.org/abs/2306.11876. arXiv:2306.11876.

Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 128–138, 2024. doi: 10.1109/wacv57701.2024.00020.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9592–9600, 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, 2022a. doi: 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 (VISAPP)*, pp. 202–213, 2022b. doi: 10.5220/0010865000003124.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ed91353f700d113e5d848c7e04a858b0-Abstract-Conference.html.

Yunkang Cao, Xiaohao Xu, and Weiming Shen. Complementary pseudo multimodal feature for point cloud anomaly detection. *Pattern Recognition*, 156:110761, 2024. doi: 10.1016/j.patcog.2024.110761. URL https://arxiv.org/abs/2303.13194. arXiv:2303.13194.

Niv Cohen, Issar Tzachor, and Yedid Hoshen. Set features for fine-grained anomaly detection. *arXiv preprint arXiv:2302.12245*, 2023. URL https://arxiv.org/abs/2302.12245.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9737–9746, 2022. doi: 10.1109/cvpr52688.2022.00951. URL https://openaccess.thecvf.com/content/CVPR2022/html/Deng_Anomaly_Detection_via_Reverse_Distillation_From_One-Class_Embedding_CVPR_2022_paper.html.

Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. SALAD – Semantics-Aware Logical Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21843–21852, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fucka_SALAD_--_Semantics-Aware_Logical_Anomaly_Detection_ICCV_2025_paper.html.

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 (AAAI)*, 2024. doi: 10.1609/aaai.v38i3.27963. URL https://arxiv.org/abs/2308.15366. arXiv:2308.15366.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.1109/cvpr52734.2025.01415. URL https://arxiv.org/abs/2412.03342. arXiv:2412.03342.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Wei Ge, Ming Tang, and Jinqiao Wang. AnomalyMoE: Towards a language-free generalist model for unified visual anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026. doi: 10.1609/aaai.v40i6.42432. URL https://arxiv.org/abs/2508.06203. arXiv:2508.06203.

Jia Guo, Shuai Lu, Lei Fan, Zelin Li, Donglin Di, Yang Song, Weihang Zhang, Wenbing Zhu, Hong Yan, Fang Chen, Huiqi Li, and Hongen Liao. One Dinomaly2 detect them all: A unified framework for full-spectrum unsupervised anomaly detection. *arXiv preprint arXiv:2510.17611*, 2025. URL https://arxiv.org/abs/2510.17611.

Yu-Hsuan Hsieh and Shang-Hong Lai. CSAD: Unsupervised component segmentation for logical anomaly detection. *arXiv preprint arXiv:2408.15628*, 2024. URL https://arxiv.org/abs/2408.15628.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/cvpr52729.2023.01878.

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.

Yejin Kwon, Daeun Moon, Youngje Oh, and Hyunsoo Yoon. LogicQA: Logical anomaly detection with vision language model generated questions. In *Annual Meeting of the Association for Computational Linguistics (ACL), Industry Track*, pp. 411–432, 2025. doi: 10.18653/v1/2025.acl-industry.29. URL https://aclanthology.org/2025.acl-industry.29/. arXiv:2503.20252.

Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. doi: 10.1109/cvpr.2010.5539872.

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*, 2025. URL https://arxiv.org/abs/2512.13671.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193. arXiv:2304.07193.

Jiahao Qiu, Xuan Qi, Tongcheng Zhang, Xinzhe Juan, Jiacheng Guo, Yifu Lu, Yimin Wang, Zixin Yao, Qihan Ren, Xun Jiang, Xing Zhou, Dongrui Liu, Ling Yang, Yue Wu, Kaixuan Huang, Shilong Liu, Hongru Wang, and Mengdi Wang. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *arXiv preprint arXiv:2505.20286*, 2025. URL https://arxiv.org/abs/2505.20286.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning (CoRL)*, 2023. URL https://arxiv.org/abs/2307.01928. arXiv:2307.01928.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *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.

Rongbin Tan, Fangfang Lin, Zhenlong Yuan, Min Qiu, Kejin Cui, Mengmeng Wang, Yi Wang, Zijian Song, Zhiyuan Wang, Jiyuan Wang, Yue Wang, Shuhan Song, and Huawei Cao. IndusAgent: Reinforcing open-vocabulary industrial anomaly detection with agentic tools. *arXiv preprint arXiv:2605.20682*, 2026. URL https://arxiv.org/abs/2605.20682.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024a. URL https://arxiv.org/abs/2305.16291. arXiv:2305.16291.

Zhiruo Wang, Graham Neubig, and Daniel Fried. TroVE: Inducing verifiable and efficient toolboxes for solving programmatic tasks. In *International Conference on Machine Learning (ICML)*, pp. 51177–51191, 2024b. URL https://proceedings.mlr.press/v235/wang24az.html.

Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelović, and Jakob Nikolas Kather. LLM agents making agent tools. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. doi: 10.18653/v1/2025.acl-long.1266. URL https://arxiv.org/abs/2502.11705. arXiv:2502.11705.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2407.10299. arXiv:2407.10299.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. doi: 10.1109/cvpr52734.2025.01416. URL https://arxiv.org/abs/2503.18325. arXiv:2503.18325.

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*, 2026. 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://arxiv.org/abs/2406.04687. arXiv:2406.04687.

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 (ECCV)*, 2022. doi: 10.1007/978-3-031-20056-4_23.

A Compiler and Runtime Details

This section records the frozen candidate families, runtime binding, BT composition, and verifier-precision weighting summarized in the main method.

Uniform CFv3 specialist compilation. The final strict-FS route applies the same compiler to every LOCO class. Each card receives the same seed 17 EfficientAD and RD4AD candidate families, five EfficientAD and three RD4AD score fields, a $1/40$ simplex grid with at least 0.5 weight on the general primitive, and the same gain and source-fold gates. Experts are mapped through empirical CDFs defined only by their same-class CFv3 controls. A specialist route is enabled only if it improves CFv3 AUROC by at least 0.50 percentage points, uses a nonzero specialist, has non-negative median source-fold gain, and is non-negative on at least three of five folds. All five cards are frozen together before LOCO evaluation. For classes with an accepted agent program, the selected semantic mixture replaces only the general primitive; the program and its previously frozen reliability weight remain unchanged.

Runtime-bound independent-image deployment. The primary ANP configuration binds those frozen cards to the implementation used in the latency experiment. For breakfast_box, the backbone runs in FP16, the normalized full DINOv2 (Oquab et al., 2024) memory bank remains FP32 (356,352 rows), and FAISS IVF-Flat uses 4,096 lists. We sweep `nprobe` in $\{2, 4, 8, 16, 32\}$ only on normal-derived CFv3 controls and freeze $\texttt{nprobe} = 4$ before reading LOCO labels. The component masks are pre-generated but score-free: they contain no feature, score, label, or decision. At inference, the active memory bank, CLIP/DINOv2 forward passes, instance matching, structural scoring, and fusion all remain online. This is the configuration used for both the LOCO accuracy row and the efficiency row.

CFv3-gated benchmark composition. The separately labeled BT configuration admits the repository AnomalyMoE score only after the same CFv3 audits. We use fixed equal ranks when they improve CFv3 AUROC; if one expert perfectly separates all controls, it remains primary and the other orders only samples in the same empirical-CDF bin via $r_{\text{primary}} + \epsilon r_{\text{secondary}}$, $\epsilon = 0.5/|\text{CFv3}|$. This cannot reverse adjacent primary bins. All decisions precede LOCO labels; BT remains necessary because AnomalyMoE normalizes over the unlabeled evaluation cohort.

Precision compilation of compatible programs. When several independently verified programs expose the same score contract, the compiler can use verifier stability as a weight rather than tuning on evaluation labels. Let a_{js} be program j's AUROC on normal-derived CFv3 trials under seed s, and let v_j be its sample variance across five seeds. With $\tilde{v}_j = \max(v_j, \epsilon)$, we freeze

$$\omega_j = \frac{\tilde{v}_j^{-1}}{\sum_k \tilde{v}_k^{-1}}, \qquad \epsilon = 10^{-12},$$

then compile $r_D(x) = \sum_j \omega_j \,\mathrm{rank}(s_j(x))$. For the video extension, shared MoE evidence is counted once: $r(x) = 0.5\,\mathrm{rank}(s_{\mathrm{MoE}}(x)) + 0.5 r_D(x)$. The variance floor, modality split, and formula are fixed; evaluation labels enter only the final metric. Evaluation-cohort ranks make this operating point FS+BT.

B Supplementary Experimental Evidence

This appendix preserves the full protocol, acceptance ledger, negative controls, calibration audit, cross-domain breakdowns, and latency diagnostics behind the condensed main-paper results. All routing and weighting decisions retain the same normal-only freeze boundary used in the main paper.

B.1 Full Protocol and Data Controls

We evaluate on four benchmark families. MVTec LOCO (Bergmann et al., 2022a) contributes five classes with both logical and structural anomalies, MVTec-AD (Bergmann et al., 2019) fifteen structural classes, VisA (Zou et al., 2022) twelve classes, and BMAD (Bao et al., 2024) the three medical benchmarks BrainMRI, LiverCT and RESC.

Synthesis, verification, selection, and calibration use normal images and the known validity of transformations applied to those images. Test labels are loaded only to compute the reported metrics. Normal images are divided into induction, validation, and calibration folds as described in Section 3; multi-seed results redraw this partition. For the matched-budget study, a total of $n \in \{30, 60, 120, 240\}$ normals per class is shared by the memory bank, program induction, validation, calibration, and the internal audit. A separate deployment-style router uses the official validation normals outside this budget and is labeled accordingly.

Our LogSAD rerun follows the repository implementation. A near-full normal coreset is built in four-image chunks, while four initial images seed the semantic histograms and instance rules, so this setting is neither a pure four-shot detector nor a four-image normal coreset. CFv3 contains 80 valid and 80 invalid controls per LOCO class. Light jitter, recoloring, same-location cloning, and small shifts preserve validity; dropping, duplicating, moving, or cross-recoloring components violates it. Each of the eight operators contributes 20 controls, all derived from training normals. The synthetic manifests and all-class selection cards are written before the LOCO evaluation manifest is opened. We report image AUROC on the all, logical and structural test subsets, the empirical FPR of the flag rule at α=0.05 on test normals, the number of online LLM calls, and per-image latency. Baselines are quoted from their original papers on the same benchmark and are marked with their protocol, whether full-shot trained, few-shot, or training-free; agent baselines are quoted on the LOCO domain of their own benchmark where stated.

B.2 LOCO Runtime Binding and Selector Diagnostics

The original configuration in Table 2 uses one primary independent-image configuration. ANP reaches class AUROCs of 95.32%, 99.36%, 95.43%, 86.97%, and 94.64%, in dataset order, and a mean of 94.34%. Its frozen counterpart without programs reaches 91.48%, so the verified program layer adds 2.87 points (10,000 class-stratified paired-bootstrap replicates, 95% CI $[+1.93, +3.84]$, $P(\Delta > 0) = 1.00$). The comparison uses the same per-image runtime binding as the latency row rather than pairing the accuracy of one route with the speed of another.

The program gain should not be read as wholly autonomous. The pushpins branch is a check-driven expert reference selected by a normal-only card; the screw_bag branch is Agent-conditioned. Removing the frozen branches one at a time in this same frozen configuration attributes $+2.10$ mean

Table 5: Component-aware context for MVTec LOCO. Values are macro equal logical/structural (L/S) AUROC, so they are not mixed with the macro pooled values in Table 2. The CSAD and SALAD rows are author-checkpoint references evaluated on the public test manifest; their upstream checkpoint or composition-map provenance was not independently retrained under validation-only selection. The ANP configurations are reported in Table 2; the complete ensemble comparison is in Appendix B.4. Bold marks the largest mean in this table, with source settings retained.

Method	Learned assets and selection	AUROC
ANP (primary)	Executable checks fitted on normal-derived data	94.56
ANP-P, cached	Fixed breakfast refinement; remaining routes unchanged	94.66
ANP-P	Aligned breakfast features and native masks	94.69
CSAD (Hsieh & Lai, 2024)	Grounded-SAM/SAM pseudo-masks; test-selected author checkpoint	95.67
SALAD (Fučka et al., 2025)	Author-provided automatic composition maps and checkpoint	96.04
ANP-E2	Retained semantic/program branches and two trained SALAD-derived detectors; shared segmentation	**96.19**

points to the expert reference (CI $[+1.43, +2.85]$) and $+0.77$ to the agent's dual-query repair (CI $[+0.12, +1.42]$). In the latter case, execution feedback changes the grounded variables, the agent writes the two semantic queries and repair logic, and the verifier selects the DINO-student head from a predefined grid. This is the direct causal evidence for automatic program generation; the broader $+2.87$-point result measures the complete verified-program layer.

The deployment binding changes only breakfast_box. The FP16 backbone still queries the full 356,352-row FP32 normalized memory bank. On CFv3, we test `nprobe` at 2, 4, 8, 16, and 32. The corresponding AUROCs are 98.88, 98.94, 98.94, 98.89, and 98.91; AP values are 98.08, 98.13, 98.12, 98.08, and 98.10. We therefore freeze `nprobe` $= 4$. Replaying all 275 breakfast_box test images then gives $95.32/97.65$ AUROC/AP with all component-cache lookups satisfied. Relative to the latest compiler before runtime binding, mean AUROC changes by only -0.07 points (CI $[-0.18, +0.02]$), which is not statistically resolved.

B.3 Precision Configuration and Repeatability

The breakfast refinement addresses a mismatch between the deployment helper and the original visual pipeline. The helper resized an integer-valued PIL image before conversion to a floating-point tensor. The original datamodule converts the source image first and then applies antialiased tensor resizing. We align both the four setup references and every query with the latter ordering. The backbone runs in FP32 and queries the unchanged FP32 memory bank with exact nearest-neighbour search. The structural response is the mean of the largest 20 entries of the averaged four-layer 64×64 anomaly map. The reference bank is retained. We first retain the existing score-free masks, then recompute the instance masks with native SAM and semantic segmentation under the aligned input pipeline. All 62 validation, 160 CFv3, and 275 test queries use freshly generated masks in the latter experiment. Neither experiment rebuilds the inherited training memory bank.

The structural and instance scores are calibrated separately using 62 normal validation images. For branch j, let u_j be its median normal score and g_{λ_j} the Yeo–Johnson transform fitted to s_j/u_j. With the mean μ_j and sample standard deviation σ_j of the transformed normals, the final score is

$$s_{\mathrm{P}}(x) = \mathrm{sigmoid}\left[\max_{j \in \{\mathrm{struct,inst}\}} \frac{g_{\lambda_j}(s_j(x)/u_j) - \mu_j}{\sigma_j}\right]. \tag{4}$$

Five source-group folds assess the calibration on 160 normal-derived CFv3 controls and the 62 unedited validation images. Each query fold uses parameters fitted on normal validation images in the other folds; the final transform is refitted on all 62 normals. The visual backbone and inherited reference bank are not refitted across these calibration folds. The resulting diagnostic AUROC is 98.11% jointly and 99.73% on CFv3 with cached masks, and 97.76% jointly and 99.05% on CFv3 with native segmentation.

The successive experiments in Table 6 are development ablations. FP32 and native preprocessing were fixed implementation checks, rather than winners selected by the earlier precision gate. Each

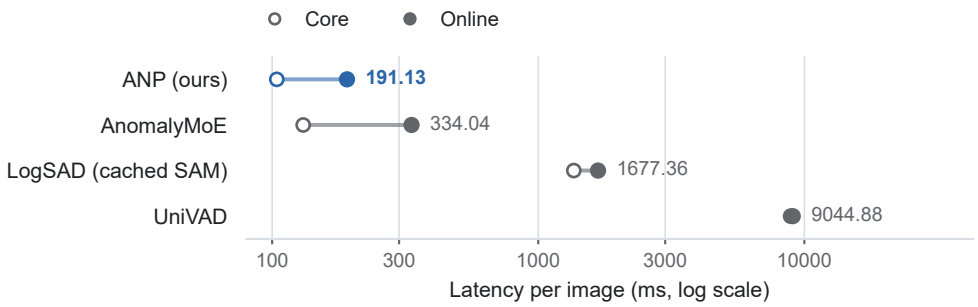

Figure 4: **Inference cost of the frozen runtime.** Open markers indicate core compute; filled markers include decoding, preprocessing, and transfer. All four methods are timed on one RTX A6000 under the protocol in Table 3. Labels give online latency in milliseconds. Model/reference setup and applicable pre-generated masks are outside both timing boundaries.

Table 6: Breakfast precision refinement on the same 275 test images. Image AUROC (%); the logical and structural subsets each include all 102 normal images. The cached-mask configuration is repeated in a fresh process; native segmentation is a separate fixed refinement.

Configuration	Pooled	Logical	Structural
Original ANP	95.32	96.79	93.97
FP16 exact retrieval, top0.5% pooling	95.52	97.15	94.02
Normal Yeo–Johnson calibration	95.64	96.87	94.51
FP32 query features	95.65	96.88	94.52
Native tensor preprocessing, cached masks	95.84	96.89	**94.87**
Independent-process replay, cached masks	95.84	96.89	**94.87**
Native tensor preprocessing and segmentation	**95.98**	**97.35**	94.71
LogSAD reproduction	95.69	97.14	94.35

configuration and its normal-fitted parameters were frozen before its test execution; the broader development process had prior LOCO exposure. Additional normal-controlled comparisons of photometric consensus, positive-evidence norms, neighbour averaging, and spatial pooling retained their reference configurations. Thus the table describes the implemented refinement rather than an independent confirmation of a newly induced program.

With cached masks, breakfast pooled AUROC is 95.84%, an improvement of 0.52 points over the original route and 0.15 over our LogSAD reference. The paired 95% bootstrap interval for the latter difference is $[-0.57, +0.93]$ points. The observed gain is numerical; the interval does not establish statistical superiority. LogSAD retains its historical four-decimal recorded scores, with no cross-label score ties. A fresh model process on another idle GPU repeats all 275 image predictions, with maximum absolute score difference zero. This verifies computation on the same images, not generalization to a new holdout.

Native segmentation raises breakfast pooled AUROC to 95.98%, with 97.35% logical and 94.71% structural AUROC. Both anomaly subsets exceed the LogSAD reference. The pooled gain over LogSAD is 0.29 points, with paired 95% interval $[-0.37, +0.99]$. The accompanying native-score control reproduces LogSAD's AUROC on all three subsets. Independently recomputing the normal calibration, test metrics, and 2,000 paired-bootstrap replicates verifies this result; all 497 new cache files contain only packed instance masks, shape, and format metadata. The lower proxy AUROC despite higher test AUROC is retained as a limitation of the proxy diagnostic. This fixed segmentation experiment has one visual run.

ANP-P applies the native-segmentation refinement to breakfast and retains the other four frozen class routes. Recomputing all 1,568 image records gives macro pooled AUROC 94.48% and macro equal logical/structural AUROC 94.69%. The cached-mask predecessor gives 94.45% and 94.66%,

respectively. The original program-layer ablation and conformal audit remain attached to their original configuration. With diagnostic feature capture, native segmentation takes 3263.60 ms mean online time (3247.73 median, 3370.21 P95), including mask generation. The cached predecessor takes 903.31 ms (900.61 median, 935.52 P95), excluding prior mask generation. Both exclude model/reference setup and are reported separately from the original runtime's latency.

A further controlled refinement replaces stepwise empirical score mappings with linear interpolation between their control-score knots. Five-fold source-group validation compares both mappings under the same expert weights for all four remaining classes. Only screw_bag passes the predefined gain and stability criteria. With all choices frozen, the five-class mean becomes 94.44%, compared with 94.45% for the cached-mask configuration; the paired 95% interval for the difference is $[-0.02, +0.01]$ points. This experiment does not support replacing the reported precision configuration.

B.4 Complementary Detector Ensemble

We test whether an independently trained component detector supplies evidence complementary to the retained semantic and program branches. The added detector follows the SALAD-derived training recipe, including normal training images, automatic composition maps, and Tiny-ImageNet auxiliary penalties. Each of two detector seeds, 42 and 43, receives exactly 70,000 updates per class; the final checkpoint is used. The segmentation model, original expert banks, program residuals, and their weights are shared. These runs therefore repeat detector training while holding the surrounding pipeline fixed.

For each class, the detector score is mapped to its empirical quantile over the same 160 CFv3 controls, producing $q_{42}(x)$ and $q_{43}(x)$. Each seed has its own reference scores. Let $u(x)$ denote the retained semantic score. The fixed complementary ensemble, ANP-E2, uses

$$m(x) = \tfrac{1}{2}u(x) + \tfrac{1}{4}q_{42}(x) + \tfrac{1}{4}q_{43}(x). \tag{5}$$

For branches with a program residual, the final score is $s_{\mathrm{E2}}(x) = s_{\mathrm{P}}(x) + w[h(m(x)) - h(u(x))]$, where $h(v) = -\log\max\{1/161, \min\{1, 1-v\}\}$ and w is the original frozen semantic weight. Other branches use $m(x)$ directly. This preserves the existing program contribution while replacing its semantic input.

The comparison also retains the single-detector ensembles and a normal-selected alternative. The latter fixes class weights to $(0, 0, 0.25, 0.5, 0)$ in dataset order, selected with five source-group folds of normal-derived controls before the second detector training began. All six combinations are frozen in advance. After training, both detectors use the same normal reference identities, and all five new references are sealed before evaluation. All 1,568 test images are retained for every route.

Table 7: **Complete detector-seed and ensemble comparison on LOCO.** AUROC (%) averaged over all five classes. Both seed runs share segmentation and historical program assets. The six fusion routes were fixed before training seed 43; no best-seed or per-class test selection is used. Bold marks column maxima.

Configuration	Pooled	Logical	Structural	Equal L/S
ANP-P	94.48	94.17	95.21	94.69
Native detector, seed 42	92.97	92.53	94.05	93.29
Native detector, seed 43	92.42	91.66	93.98	92.82
Normal-selected, seed 42	95.62	95.28	96.35	95.82
Fixed equal, seed 42	95.89	95.90	96.26	96.08
Normal-selected, seed 43	95.55	95.15	96.36	95.76
Fixed equal, seed 43	95.95	95.78	**96.53**	96.16
Normal-selected, two seeds	95.62	95.27	96.39	95.83
ANP-E2 (fixed equal, two seeds)	**96.00**	**95.97**	96.41	**96.19**

ANP-E2 reaches 96.00 macro pooled and 96.19 macro equal L/S AUROC. Its equal L/S mean is 0.15 points above the SALAD author-asset reference and 0.52 points above CSAD. The paired 95% interval for the SALAD difference is $[-0.79, +1.05]$ points, based on 2,000 within-class subtype-stratified bootstrap replicates. The reference methods retain their original training and asset prove-

nance. This supports the table's numerical ordering without establishing a statistically resolved advantage under matched training conditions. All six routes, including the lower-performing normal-selected alternatives, are included in Table 7.

The fixed two-detector ensemble improves pooled AUROC by 0.11 points over the corresponding seed-42 ensemble (95% CI $[-0.07, +0.28]$). On breakfast, its 95.05 is below ANP-P's 95.98; the reported ensemble mean uses its complete five-class route. This is a development experiment with prior LOCO exposure and shared segmentation and historical programs. It evaluates detector complementarity; the following intervention measures its retained program terms. The autonomous generation study retains its own configuration. The original accuracy evaluation uses cached composition maps. We subsequently execute the complete breakfast pipeline from RGB, generating both LogSAD/SAM and composition masks online before running the two detector seeds and fixed fusion. Five validation normals warm up the runtime; all 275 test images are then measured once. Mean, median and 95th-percentile latency are 3701.77, 3682.30 and 3879.21 ms, respectively, on one RTX A6000. Model and reference setup takes 39.12 s and is excluded from steady-state latency. The run retains diagnostic map-capture overhead and has no query-cache hits. All composition masks and score ranks match the accuracy execution; maximum E2 score difference is 1.2×10^{-16}. This measurement covers breakfast rather than a five-class latency average.

B.4.1 Program interventions in the ensemble

The ensemble permits a direct intervention on the retained explicit program terms. For pushpins and screw_bag, its score can be written as $s_{\mathrm{E2}}(x) = wh(m(x)) + e_\pi(x)$, where e_π is the frozen program evidence. We set $e_\pi = 0$ for both branches together and for each branch separately. All detector scores, reference distributions, fusion weights and clipping rules remain fixed. The other three class routes are unchanged. Full scores and the two inherited program terms are reconstructed from their source arrays before computing the interventions.

Table 8: **Explicit program branches in the fixed ANP-E2 configuration.** Five-class mean AUROC (%). Each intervention sets the indicated additive program evidence to zero while retaining both detector seeds, normal references, semantic weights and score clipping.

Configuration	Pooled	Logical	Structural	Equal L/S
ANP-E2	96.00	95.97	96.41	96.19
Without both program branches	95.45	94.71	96.84	95.77
Without pushpins expert reference	95.65	95.24	96.48	95.86
Without screw_bag agent program	95.80	95.43	96.77	96.10

Retaining both program branches contributes 0.55 pooled AUROC points (95% paired CI $[-0.05, +1.21]$); the corresponding equal-L/S contribution is 0.42 points (CI $[-0.16, +1.02]$). The separate pooled contributions are 0.35 for the pushpins expert reference and 0.20 for the screw_bag agent program. These are point-estimate gains with intervals spanning zero. The study measures existing program terms within the stronger ensemble; the autonomous generation study evaluates a different question. An independent audit checks all 1,568 image records and replays 12,000 bootstrap values using weighted pair concordance with explicit treatment of score ties.

B.5 Additional Selector Diagnostics

The specialist selector is also stable beyond the original two seeds. With the route fixed, EfficientAD seeds 17, 29, 43, 71, and 101 give juice AUROC $(99.22 \pm 0.05)\%$ (sample s.d.; range 99.17%–99.28%). Ten independent 80% CFv3 source-group selectors choose seed 17 in every split, although they divide between $1 - p$ (six splits) and `map_st_max` (four). Three complete $70/30$ reference/calibration repeats retain the same cards; their all-anomaly AUROC has sample s.d. 0.15 percentage points and range 92.87%–93.14% for the uniform-specialist ablation.

For completeness, the separately tested cohort-transductive compiler reaches 93.83%, about 0.51 points below ANP. Exposing the same BT expert pool to all classes lowers that mean further to 92.37%. We keep both as protocol diagnostics rather than as headline rows. The lightweight ANP-fast route likewise remains a speed ablation: it reaches 81.50% without an external semantic detector and satisfies the predeclared 2α budget, but it is not used for the primary latency claim.

B.6 Grounding Source Matters More than Grounding Accuracy

The comparison yields two reproducible observations. First, a GT support mask is not necessarily the right representation for a check. On splicing it covers the whole harness, whereas the program needs the separate connector–cable–connector relation. Check-driven decomposition improves logical AUROC by 9.00 percentage points and reaches $(87.60 \pm 0.30)\%$ over five splits at FPR $(3.70 \pm 0.90)\%$. Second, choosing a more specific target can still fail. The hand-written screw_bag mask captures the saturated bag instead of the screws and falls to 47.70%, below the GT-derived mask. This failure motivates testing the resulting measurement, not accepting a grounding from its visual plausibility alone.

B.7 Normal-Derived Verification Controls

The candidate comparison includes expert and agent screw_bag programs with acceptance/downstream scores $0.87/0.48$ and $0.81/0.48$, respectively. Both pass the proxy task with limited downstream discrimination. The controls do not reproduce the fine length and thread differences relevant to this class; the resulting mismatch may also involve the shared perceptual measurements. This motivates checking grammar coverage and measurement sensitivity together.

Verification helps most when the program exposes the relevant relation (Table 10). All six counting/relation cells improve across expert-written, GT-derived, and agent-generated groundings, by 0.10 to 0.32 all-AUROC (mean $+0.20$). Breakfast_box and juice_bottle are mixed, including a -0.11 change for the breakfast oracle. On screw_bag, all three grounding sources are negative because the synthetic proxy favors programs that do not match the real defect. Verification therefore amplifies the quality of the proxy; it cannot repair a missing anomaly mode. The procedure itself remains label-free: it selects the pin-head segmentation with count CV 0 and rejects a naive blob alternative with count standard deviation 12.8.

Table 11 summarises the five mechanism tests cited from the main paper. Each row states its own metric and controlled comparison, and the effects are not pooled across rows; the fuller per-mechanism evidence appears in the other subsections of this appendix.

Table 11: Mechanism tests. Each row states its own metric and controlled comparison; effects are not pooled across rows.

Question	Controlled evidence
Does the grounding target matter?	On splicing logical AUROC, GT-derived object support gives 0.78 and the measurement-matched connector–cable–connector decomposition gives 0.88 ($+9.2$ points).
Does discriminative verification help?	Across six counting/relation cells and three grounding sources, verified programs improve all-anomaly AUROC by $+9.6$ to $+31.6$ points (mean $+19.7$).
Can feedback repair synthesis?	On splicing, replacing count stability with the discriminative gate changes the accepted artifact from 0.66 to 0.76 ± 0.02 logical AUROC.
Does an agent-written repair affect the system?	Removing the screw_bag semantic repair lowers mean LOCO AUROC by 0.77 points, CI $[+0.12, +1.42]$.
Where does the proxy fail?	A screw_bag candidate passes synthetic discrimination at 0.81 but reaches 0.48 downstream. A similar result with expert grounding motivates examining both grammar coverage and shared perceptual measurements.

B.8 Autonomous Synthesis with the Discriminative Gate

The count-stability gate accepts $6/9$ class attempts, but its splicing program shows why that criterion is incomplete. Two color-stable connector blobs have CV 0.06 and pass the gate, yet the program reaches only 0.66 logical AUROC because it omits the relation that defects actually change.

The discriminative gate makes a different decision. On splicing it accepts the second candidate despite count CV $0.26 > 0.10$, because the regions reach label-free discriminative AUC 0.90. With that grounding held fixed across five splits, automatic selection obtains $(76.40 \pm 2.40)\%$ logical AUROC at FPR $(3.90 \pm 1.40)\%$. This is 10.70 percentage points above the count-gate artifact, although the expert-written reference remains stronger at $(87.60 \pm 0.30)\%$.

Screw_bag reveals the remaining limitation. In the first run, the only executable candidate produces 43 reasonably stable regions but discriminative AUC 0.25, so the new gate rejects it. Long repair

Table 9: Grounding-source probe on LOCO under an identical induction pipeline (same programs, folds, α); only the grounding source changes. Check-driven masks here are *expert-written* references (labeled), used to isolate the target-design question from synthesis quality. Logical AUROC.

Grounding source	Juice	Push.	Screw	Splic.
Generic prompt (GroundingDINO+SAM)	—	0.57†	—	0.67†
GT-derived support mask	0.77	0.91	0.66	0.78
Check-driven (expert-written)	**0.81**	0.85	0.48	**0.88**

†all-anomaly AUROC (logical subset not reported in that run).

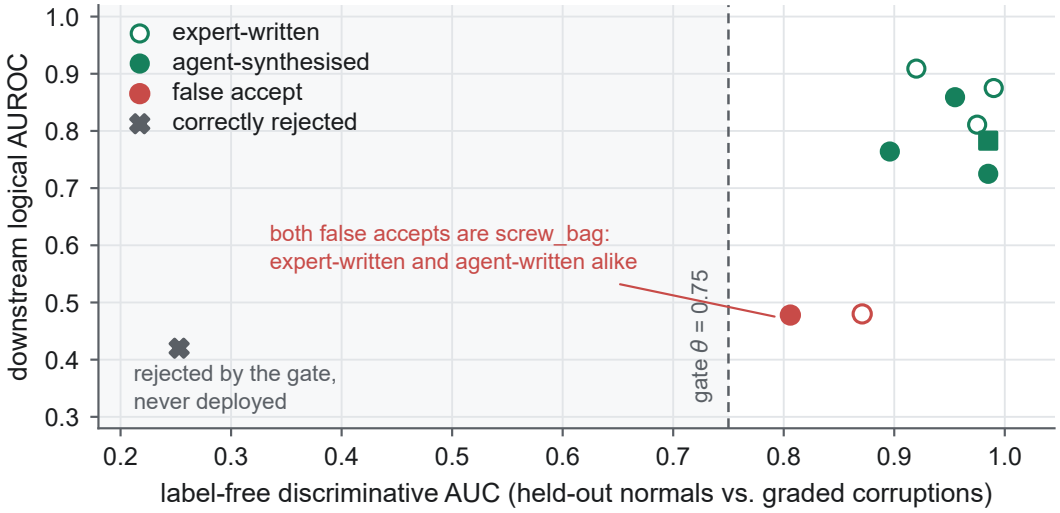

Figure 5: **Acceptance score and downstream detection.** Each point is a candidate program. The horizontal axis is its normal-derived discriminative score and the vertical axis its downstream logical AUROC. Marker styles distinguish grounding sources. The screw_bag candidates illustrate the difference between proxy acceptance and real-defect sensitivity.

histories caused the other attempts to stop compiling; restarting each attempt from a short "fresh-draft" context removes that failure. In the second run, the gate rejects two further candidates at 0.59 and 0.68, including one with perfect count CV, before accepting a candidate at 0.81. That accepted program still reaches only 0.48 downstream AUROC. The same mismatch occurs for the expert-written grounding (0.87 → 0.48 in Fig. 5), which points to the corruption grammar rather than the code author: component-level perturbations do not represent abnormal screw lengths inside the bag. We use this execution trace to ask the agent for a different set of grounded variables.

Execution-driven semantic repair. Given the normal images and the failed measurements, the repair agent proposes two semantic queries, `threaded_shafts` and `ring_hardware`. It infers the normal typed multiset, two shafts and four rings, excludes the packaging strip, and writes a Hungarian-matching plus kNN set scorer. The same specification is then compiled into a higher-resolution DINO component student. The verifier chooses among these candidates using held-out-normal FPR and counterfactual missing, extra, type, and geometry changes; no LOCO anomaly is supplied. Table 13 shows the resulting progression.

The selected student reaches 70.65%/72.12% all/logical AUROC. An after-the-fact label oracle over the same generated candidates reaches only 71.21%/72.72%, so normal-only selection captures most of the available standalone headroom. The program is more valuable as complementary evidence: reliability-shrunk compilation reaches 87.64% all AUROC in this study. In the current frozen detector configuration, its removal costs 0.77 mean AUROC percentage points (Section B.2). The semantic queries, typed normal specification, matching logic, and executable scorer all originate in the repair loop; only the student head is chosen from a predefined grid.

Table 10: Verification ablation across all LOCO classes and three grounding sources (expert-written, GT-derived "oracle," agent-synthesized). Δ = verified − no-verification, all-anomaly AUROC.

Class	Grounding	Verified / No-verif. (all AUROC)			Δ
pushpins	expert / oracle / agent	0.91/0.59	0.91/0.61	0.80/0.61	+0.32 / +0.30 / +0.19
splicing	expert / oracle / agent	0.82/0.67	0.74/0.64	0.71/0.58	+0.15 / +0.10 / +0.13
juice	expert / oracle / agent	0.73/0.69	0.71/0.74	0.67/0.68	+0.04 / −0.03 / −0.01
breakfast	oracle / agent	0.68/0.79	0.74/0.74		−0.11 / 0.00
screw_bag	expert / oracle / agent	0.46/0.56	0.66/0.69	0.49/0.58	−0.10 / −0.04 / −0.09

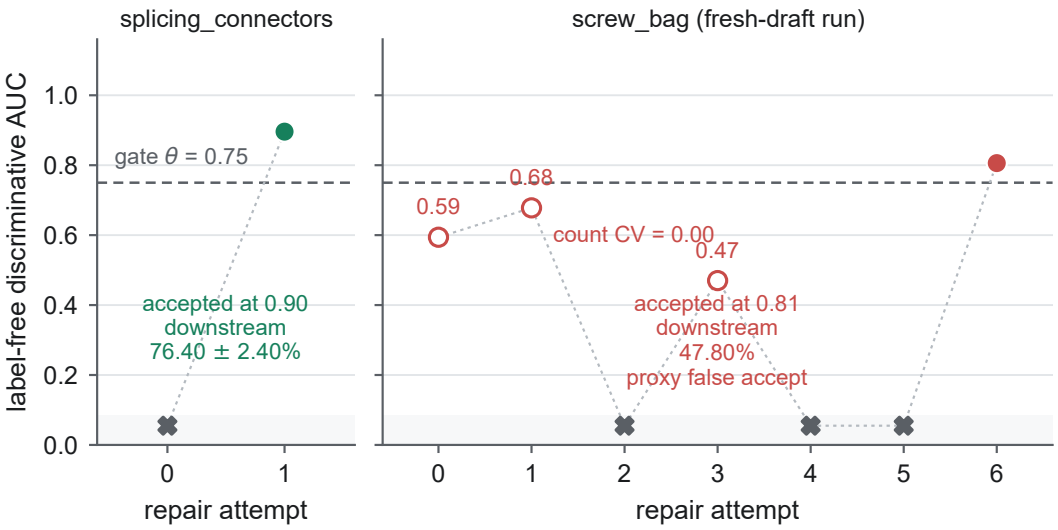

Figure 6: **Program generation and verification trajectories.** The splicing episode reaches an accepted program after execution feedback. The screw_bag episode includes non-executable candidates and gate rejections before a proxy-accepted program. The candidate with count CV 0.00 is rejected at discriminative score 0.68, illustrating the additional information provided by the control task over normal-output stability.

Synthesis stability. Across three proposer seeds, all five LOCO classes, and four VisA classes, the loop completes 27/27 class–seed runs and accepts 16/27 (59.3%; Wilson 95% CI 40.7–75.5%). Breakfast and juice are accepted in all three seeds, pushpins and screw_bag in none, and splicing in one; on VisA, cashew, pcb1, and pipe_fryum are accepted in all three, while capsules is never accepted. The final failures consist of six compile errors, one execution error, and four low-discrimination rejections.

Acceptance measures whether the artifact satisfies the normal-derived tests, not whether it beats a strong visual detector. Under one common downstream protocol, none of the 16 accepted artifacts surpasses its structural branch. Mean all-anomaly AUROC is 0.65/0.57/0.67 on LOCO breakfast/juice/splicing and 0.87/0.81/0.72 on VisA cashew/pcb1/pipe_fryum, with per-class gaps of −0.11 to −0.30. Discriminative AUC is only weakly correlated with downstream AUROC (Pearson r=0.32, Spearman ρ=0.31). These runs diagnose the generic corruption grammar; the targeted semantic repair above is a new feedback policy and is therefore reported separately.

Primitive-family removal. Removing one template family before verification leaves the sparse selector with programs on at most 3/5 classes. Count, geometry, and sequence removal change neither coverage nor system AUROC, while removing topology changes mean AUROC by +0.12 points. Without color, coverage falls to 2/5 but system AUROC rises from 0.66 to 0.72 because the affected class returns to a stronger structural detector. The gain therefore comes from verified routing and fallback behavior, not from exposing the compiler to more template names.

Table 12: Autonomous generate–verify–repair synthesis: complete ledger, both gates, all attempted classes. "disc. AUC" is the label-free discriminative signal at acceptance; detection numbers are after automatic FPR-feasible selection and conformal calibration. Failure rows are part of the result.

Class	Proposer	Gate	Outcome	disc. AUC	Logical AUROC	FPR@0.05
pushpins (LOCO)	text	count-stab.	accepted (3 att.)	—	0.86–0.86	0.12
juice_bottle	mm	count-stab.	accepted (2 att.)	—	0.72	0.00
breakfast_box	mm	count-stab.	accepted (3 att.)	—	**0.86**	**0.03**
cashew (VisA)	text	count-stab.	accepted (1 att.)	—	0.92–0.94[a]	0.02–0.04
pipe_fryum (VisA)	mm	count-stab.	accepted (1 att.)	—	0.78[a]	0.00
pcb1 (VisA)	mm	count-stab.	accepted (3 att.)	—	0.74[a]	0.01
splicing_conn.	mm	count-stab.	*accepted but uninformative*	—	0.66	0.05
screw_bag	mm	count-stab.	✗ (4 att.)	—	—	—
capsules (VisA)	mm	count-stab.	✗ (9 att., CV 0.13)	—	—	—
splicing_conn.	mm+sem.	discriminative	accepted (2 att.)	0.90	0.76±0.02[c]	**0.04±0.01**
screw_bag (run 1)	mm+sem.	discriminative	✗ (7 att.)[b]	0.25	—	—
screw_bag (run 2, fresh-draft)	mm+sem.	discriminative	accepted (7 att.), *proxy false accept*	0.81	0.48	0.03

[a] VisA has no logical/structural split; the number is all-anomaly AUROC.
[b] one executable candidate was correctly *rejected* by the gate (disc. AUC 0.25); the remaining attempts failed compilation (proposer degradation over long repair contexts).
[c] mean±std over five induction/validation/calibration split seeds with the accepted grounding fixed.

Table 13: Normal-feedback repair on screw_bag. All candidates are selected without LOCO test labels; "generic" rows are failed diagnostic predecessors, and the last row is the frozen reliability-shrunk accuracy route. Its latency is the sum of separately measured LogSAD and Agent-student cores, not a directly timed chain; all rows exclude pre-generated mask construction.

Compiled artifact	All AUROC	Logical AUROC	Online core (ms)
Generic mask crop-DINO	62.30	60.67	1008.6
Generic GroundingDINO component program	65.80	61.36	241.6
Agent two-query geometry program	64.69	68.72	—
Agent two-query high-resolution DINO student	70.65	72.12	159.60
ANP, reliability-shrunk with general primitive	**87.64**	**85.29**	3554.91

CFv3 independence and sensitivity. Ten deterministic 80% source-group splits test whether one CFv3 sample determines the compiler. Refrozen fields and weights vary, but mean all-anomaly AUROC remains $(92.96 \pm 0.11)\%$ (range 92.72%–93.11%), close to 92.99% with full CFv3. Breakfast is disabled in every split; juice, pushpins, and screw_bag are enabled in every split, and splicing in nine. Leaving out cardinality, spatial, or appearance controls changes mean AUROC by $+0.03/-0.12/-1.01$ points, respectively. Appearance removal also disables juice and splicing, identifying which part of the proxy the result depends on.

Gate threshold. We replay every executable synthesis trace to measure the effect of θ without new model calls. At θ=0.60, another screw_bag candidate is admitted but reaches only 53.95% logical AUROC. At θ=0.90, every candidate is rejected, including the useful splicing grounding (89.61% discriminative AUC; 74.81% logical AUROC). The chosen value, 0.75, keeps that splicing program and one screw_bag false accept. Across the recorded screw_bag candidates, discriminative AUC ranges from 25.33% to 80.56%, while downstream all-anomaly AUROC remains between 44.73% and 58.09%. This pattern is consistent with a class-specific proxy mismatch.

Strict normal-image budgets. We count the union of image identities used for detector fitting, program induction, selection and calibration against a total per-class budget $n \in \{30, 60, 120, 240\}$. Five classes and three coupled seeds produce 60 complete runs. The joint selector compares ordinary features and generated programs using normal-derived utility, with fixed tie breaking; every comparator receives the same available observations. All seven complete routes are reported in Table 14.

The joint automatic route reaches 78.02, 79.52, 80.44 and 80.82% as the budget increases. It leads at $n = 120$ and 240 and ties the feature-based controls at $n = 60$. At $n = 240$, its improvement over feature selection is $+0.16$ points, with a within-budget simultaneous 95% interval of $[+0.05, +0.28]$

Table 14: Strict total-normal-budget comparison on all five LOCO classes. Values are mean equal logical/structural image AUROC (%) and population standard deviation over three coupled split/training seeds. Every route uses the same total per-class image budget; bold marks exact column leaders.

Route	$n = 30$	$n = 60$	$n = 120$	$n = 240$
Structural detector	78.03 ± 0.51	79.28 ± 0.19	80.10 ± 0.01	80.48 ± 0.19
Fixed program	78.03 ± 0.51	79.28 ± 0.19	80.07 ± 0.06	80.47 ± 0.19
Feature selection	**78.06** ± 0.54	**79.52** ± 0.34	80.35 ± 0.29	80.66 ± 0.33
First program	77.99 ± 0.51	79.39 ± 0.23	80.34 ± 0.29	80.80 ± 0.32
Program-only automatic	77.99 ± 0.51	79.39 ± 0.23	80.34 ± 0.29	80.79 ± 0.31
First program + features	78.02 ± 0.55	**79.52** ± 0.34	80.35 ± 0.28	80.76 ± 0.25
Joint automatic	78.02 ± 0.55	**79.52** ± 0.34	**80.44** ± 0.38	**80.82** ± 0.29

across five contrasts. Its difference from the first-program-plus-features control is $+0.06$ points ($[-0.03, +0.14]$), so this experiment does not separately resolve the benefit of feedback. The four budgets select nonzero generated programs in $1/0/2/3$ of their 15 class–seed runs; the $n = 60$ tie comes from ordinary features. These comparisons use one complete frozen selection rule across all budgets, rather than the best result from different versions.

Calibration remains a separate question. At $n = 30$ and 60, finite calibration resolution prevents rejection at $\alpha = 0.05$. Mean test-normal flag rates at $n = 120$ and 240 are 10.51% and 15.34%. The ranking gains therefore do not establish nominal false-positive control. Intervals condition on the trained models and coupled seeds and are not adjusted jointly across the four budgets.

B.9 Calibration and Deployment-Normal Audits

Exchangeability is not automatic in a deployed benchmark. For pushpins, the structural branch estimates FPR 3.30% on a held-out training fold but flags 58.00% of test normals. Their scores are systematically higher (median 0.48 versus 0.42 in calibration), an acquisition shift that another training split cannot reveal. The logical branch shows a milder version of the same effect ($(13.60 \pm 5.80)\%$ FPR across seeds), whereas splicing remains at $(3.70 \pm 0.90)\%$.

Our deployment protocol has three stages. It first calibrates on training normals, then measures the flag rate on a second, deployment-side normal batch. If this audit exceeds 2α, the branch is recalibrated on deployment-distribution normals. A branch that still exceeds the budget on a fresh normal batch is marked untrusted for that class. This procedure uses no anomaly labels.

Juice_bottle, screw_bag, and splicing pass their audits at $0.02/0.05/0.03$ and retain training calibration, yielding test FPR $0.01/0.07/0.06$. On splicing, unnecessary recalibration would have raised FPR from 0.06 to 0.17, so the pass decision matters. Breakfast_box alarms at 0.18 and is repaired from 0.10 to 0.01. Pushpins alarms at 0.26, but recalibration only changes 0.58 to 0.47 because the shift continues; the structural branch is therefore rejected. Its verified logical fallback is recalibrated on 69 official validation normals after being fitted with disjoint induction/selection/calibration folds (186/74/112 images). The fallback lowers test-normal FPR from 0.12 to 0.04 and changes all-anomaly AUROC from 0.80 to 0.80. All five routed LOCO operating points then meet the predeclared 2α budget, and three meet α.

The audit across domains. The same rule provides both a sensitivity and a specificity check across datasets. On VisA, a second same-distribution normal fold leaves $11/12$ classes untouched; cashew produces one small-sample false alarm at 0.11 and is harmlessly recalibrated. BMAD separates patients between train and test. Using one half of deployment normals for audit and the other for evaluation, the rule alarms on LiverCT (0.30) and RESC (0.25), repairing FPR from 0.29 to 0.05 and from 0.27 to 0.05. BrainMRI falls in the gap between α and the 2α alarm threshold: an audit rate of 0.08 does not trigger, and the eventual 0.10 operating point is marked untrusted. Across 20 class decisions on LOCO, VisA, and BMAD, there are 18 correct calls, one harmless false alarm, and this one gray-zone miss.

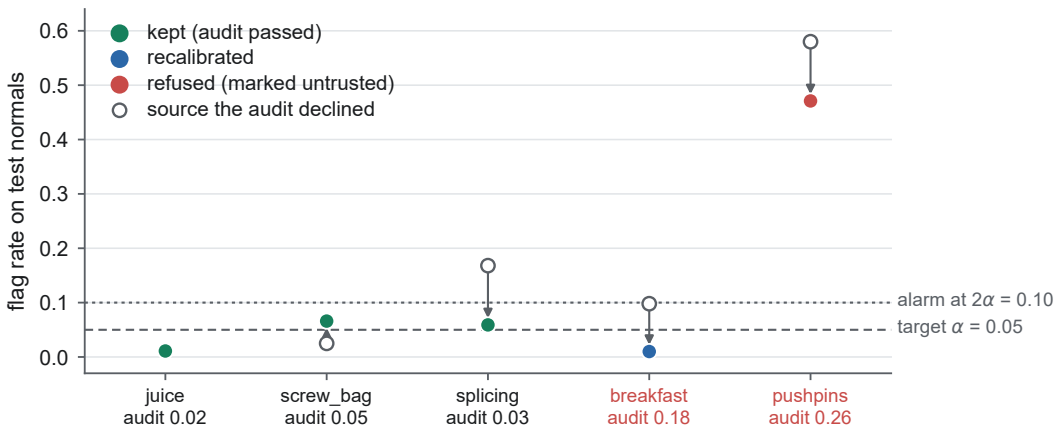

Figure 7: **Normal-only calibration audit.** The deployment-normal flag rate is checked at target $\alpha = 0.05$. The audit retains the training calibration for three classes, triggers recalibration for breakfast_box, and identifies an untrusted pushpins branch. Source-specific test rates and the separate routed fallback are detailed below.

B.10 Cross-Domain Transfer

VisA split verification. The legacy VisA directory contains all 9,621 normal images in its normal pool, including the 962 normals also present in the test directory. Identical-image hashes confirm this overlap. We exclude the resulting VisA scores from the comparative tables. The replacement uses the official `1cls.csv` split: 8,659 training normals and 2,162 test images, with zero training–test hash overlap. Two models are initialized independently; calibration uses held-out training normals and evaluation uses the final checkpoint. VisA diagnostic values elsewhere in this appendix refer to the legacy pool and are not evidence of performance on a disjoint test set.

The previously circulated 98.73% VisA value was produced by a first-party fusion over that legacy pool. It is retained in the experiment archive for traceability, but is not a valid held-out benchmark number because 962 test normal images are duplicated in the normal pool. It is therefore excluded from the comparative tables; the main table uses the complete official-split composition.

The official-split memory study uses 7,795 training normals to construct reference banks and 864 disjoint training normals for calibration. A frozen DINOv2-B/14 encoder with register tokens processes each RGB image at 392×392 resolution. We average its last four normalized token layers and normalize each resulting vector to unit length. The patch bank samples 24 tokens per training image and caps each class at 16,384 tokens; the global bank retains every training class token. Image scores are the mean of the largest 1% of nearest-patch squared distances and the nearest global squared distance, respectively. Separate normal-fitted Yeo–Johnson transforms map both scores to $\log(1+\exp(z))$ evidence, whose equal mean defines the composition control in Table 15. The banks, calibration and composition rule are fixed before test scoring; the comparison uses the complete official test split with no image-hash overlap.

As a complementary local-texture control, we extract ImageNet-pretrained WideResNet-50-2 layer-2 and layer-3 features from complete 448×448 images, apply 3×3 average pooling and concatenate aligned features. Each class retains 8,192 farthest-first and 8,192 random normal patches; selection uses a fixed 128-dimensional random projection, while scoring uses Euclidean distance in the original 1,536-dimensional feature space. The predeclared image readouts are the maximum, top 0.1% mean (primary), and top 1% mean. To avoid finite-precision saturation of a power transform, we standardize $\log(1+s/u)$ on held-out normals, where u is their median score, and apply $\log(1+\exp(z))$. This fixed rule preserves every within-class raw-score rank and yields 94.47% for the primary readout. All three whole-dataset controls appear in Table 15; they measure local expert quality separately from reconstruction composition.

The VisA overview reports the strongest complete primary configuration in this development comparison. All twelve classes and both complete seed models use each experiment's fixed composition rule. Component and alternative readouts are reported separately below.

Table 15: **Frozen memory controls on the official VisA split.** Mean image AUROC and AP (%) over all 12 classes and 2,162 test images. All controls use the same 7,795 fit and 864 calibration images. The local texture block uses a WideResNet encoder and stable normal-only log calibration.

Configuration	AUROC	AP
DINO patch memory	95.24	95.80
DINO global memory	89.88	90.57
DINO equal memory	95.53	95.43
Local texture, maximum	94.38	95.76
Local texture, top 0.1%	94.47	95.69
Local texture, top 1%	93.57	94.50

We also refit both complete reconstruction models on the full official normal-training pool, retaining a fixed 10,000-step cosine-decay schedule. The native scores of each final model are standardized using its training-normal log-score distribution, followed by softplus and equal averaging. These are in-sample normal score scales, rather than held-out calibration. The configuration tests normal coverage and initialization variance; it is an additional trained reconstruction ensemble, without a new program-generation stage.

Table 16: **Full-normal reconstruction ensemble on VisA.** Mean image AUROC and AP (%) over all 12 classes and 2,162 official test images. Each refit adds 10,000 fixed cosine-decay updates to a 40,000-update model, using all 8,659 official training normals. Both complete models have equal weight.

Configuration	AUROC	AP
Original two-model ensemble	99.31	99.42
Full-normal refit, seed 1	99.25	99.38
Full-normal refit, seed 2	99.39	99.50
Full-normal equal ensemble (primary)	99.35	99.47

Table 17: **Final reconstruction refits on the official VisA split.** Mean image AUROC and AP (%) over all 12 classes. Each refit adds 10,000 fixed cosine-decay updates to seed 1. Held-out refits preserve the 7,795/864 fit/calibration split; full-normal refits fit all 8,659 official training normals and use the native score without held-out calibration.

Configuration	AUROC	AP
Original reconstruction	99.09	99.22
Held-out refit	99.23	99.35
Held-out refit, calibrated	99.23	99.35
Held-out refit, layer composition	99.27	99.37
Original + refit, equal ensemble	99.18	99.30
Full-normal refit	99.25	99.38

Full-image and overlapping-tile inference preserve border evidence and increase the scale of small defects. Each view uses the same final model and its own normal-fitted score scale, followed by equal evidence averaging.

We also estimate the normal reconstruction mean and variance at each image location using fitting-normal images. A fixed 10% variance shrinkage toward the spatial average regularizes the standardized residual. Independent normal calibration then combines native and residual evidence equally. The complete controls appear in Table 19.

To retain small-defect evidence, we pool both the highest 1% and 0.1% of each native- and high-resolution anomaly map. Normal-only score scales allow the four readouts and the two complete seeds to contribute equally. Table 20 reports the fixed readouts and memory controls.

Table 18: **Full-image and overlapping-tile inference on VisA.** Mean image AUROC (%) over all 12 classes and 2,162 official test images. Each fixed final model retains its normal calibration set. Both complete seed models receive equal weight; no additional model training is used.

Configuration	Seed 1	Seed 2	Equal ensemble
Native view	99.09	99.39	99.31
Full-image view	98.11	98.41	98.37
Overlapping tiles	94.51	95.22	95.34
Full + tiled (primary)	98.06	98.54	98.49

Table 19: **Normal spatial residual calibration on VisA.** Mean image AUROC and AP (%) across all classes. Spatial means and regularized variances use only training-normal images. Native and standardized residual evidence receive equal weight, calibrated on disjoint normal images. The two complete seed models receive equal weight under FS.

Configuration	AUROC	AP
Native reconstruction	99.31	99.42
Spatially standardized reconstruction	99.07	99.16
Native + spatial reconstruction	99.29	99.39
Memory + native	99.13	99.21
Memory + spatial residual	98.91	98.98
Memory + native/spatial (primary)	99.06	99.14

Table 20: **Multi-scale and multi-area readout on VisA.** Mean image AUROC and AP (%) across all classes. The same final model uses 392- and 518-pixel crops and pools the highest 1% and 0.1% of each anomaly map. The four normal-calibrated scores receive equal weight. The two complete seed models receive equal weight under FS.

Configuration	AUROC	AP
Native, top 1%	99.31	99.42
High resolution, top 1%	99.27	99.43
Two scales, top 1%	99.41	99.53
Two scales, top 0.1%	99.29	99.42
Two scales and two areas (primary)	99.47	99.57
Memory + native	99.13	99.21
Memory + high resolution	99.26	99.38
Memory + two scales	99.29	99.38
Memory + small area	99.40	99.47
Memory + two scales and two areas	99.39	99.47

Table 21: **Accuracy and measured latency of the VisA scale/area composition.** Both final models reside on one GPU and run sequentially at batch size one. We time all 2,162 official test images after 24 normal-image warmups, including image decoding, preprocessing, all model forwards, pooling and calibrated fusion. Model loading and audit calculations are excluded. Accuracy is recomputed in this timed run. This measures the VisA configuration separately from the LOCO timing study.

Configuration	AUROC / AP (%)	Mean (ms)	95th pct. (ms)
Two-model scale/area composition	99.47 / 99.57	327.11	329.90

NVIDIA RTX A6000; tail latency is the 95th percentile.

We increase the reconstruction input from 392 to 518 pixels while preserving the resize-to-crop ratio and orientation. Each scale uses the same held-out normal calibration, and their evidence receives equal weight. The complete composition and its fixed controls appear in Table 22.

The same local-memory construction is also evaluated on MVTec 3D-AD, using the plane-removed RGB image and the HOT-coloured depth image as separate observations. Memory construction uses 2,394 fit images and 262 held-out normal images for calibration. At the 56×56 feature grid, cells

Table 22: **Full-context two-scale inference on VisA.** Mean image AUROC and AP (%) across all classes. The same final model uses 392- and 518-pixel crops with a fixed field of view; the two calibrated scores receive equal weight. The two complete seed models receive equal weight under FS.

Configuration	AUROC	AP
Native reconstruction	99.31	99.42
High-resolution reconstruction	99.27	99.43
Two-scale reconstruction	99.41	99.53
Memory + native	99.13	99.21
Memory + high resolution	99.26	99.38
Memory + two scales (primary)	99.29	99.38

Table 23: **DINO-memory composition on the official VisA split.** Mean image AUROC and AP (%) over all 12 classes and 2,162 test images. Both complete reconstruction seeds receive equal weight. Patch, global and reconstruction evidence are calibrated on held-out normal images under FS.

Configuration	AUROC	AP
Reconstruction, equal ensemble	99.31	99.42
Patch memory + reconstruction	98.85	98.90
Patch + global + reconstruction (primary)	98.37	97.97

Table 24: **Reconstruction experts on the official VisA split.** Mean image AUROC and AP (%) across all 12 classes and 2,162 test images. Each fixed final model uses 40,000 updates, 7,795 fitting-normal images and 864 disjoint calibration-normal images. These are individual expert controls.

Readout	Seed	AUROC	AP
Native reconstruction	1	99.09	99.22
Continuous layer calibration	1	99.13	99.25
Empirical layer calibration	1	98.67	98.35
Native reconstruction	2	99.39	99.49
Continuous layer calibration	2	99.40	99.50
Empirical layer calibration	2	98.88	98.67

Table 25: **Reconstruction seed ensembles on the official VisA split.** Mean image AUROC and AP (%) over all 12 classes and 2,162 test images. Each readout averages the two complete 40,000-update models after the predeclared normal-only calibration. No additional training is used.

Configuration	AUROC	AP
Native, normal-standardized ensemble	99.31	99.42
Continuous layer-tail ensemble	99.35	99.45
Empirical layer-tail ensemble	99.04	98.91

with less than 50% valid-depth occupancy are excluded in both streams. This mask depends only on observed depth. The fixed equal RGB–depth composition reaches 92.38% AUROC with the predeclared top-0.1% readout; the component scores are 86.27% and 86.82%. Table 28 retains the maximum and top-1% controls. These results isolate local appearance and depth evidence before combination with a learned reconstruction expert; they do not replace the complete configurations in Table 1.

Table 1 gives the six-benchmark overview in the main paper. Table 29 reports paired composition gains. The published overview rows are transcribed from AnomalyMoE's Table 1 (Gu et al., 2026), whose baselines were reproduced in its unified, single-model setting; they are not our reruns. The BMAD column averages BrainMRI, LiverCT, and RESC with equal weight. The additional Dinomaly2-L reference uses its Tables 2 and 4 (Guo et al., 2025). Our overview row uses dataset-specific compiled configurations, including ANP-P for LOCO, the fixed RGB layers $\{5, 8, 11\}$ top-

Table 26: **Fixed composition on the official VisA split.** Mean image AUROC and AP (%) across all 12 classes and 2,162 test images. Both complete reconstruction seeds receive equal weight; local memory and reconstruction then contribute equally under FS, with normal-only log calibration.

Configuration	AUROC	AP
Reconstruction, seed 1	99.09	99.22
Reconstruction, seed 2	99.39	99.49
Reconstruction, equal ensemble	99.31	99.42
Local memory, maximum	94.38	95.76
Local memory, top 0.1%	94.47	95.69
Local memory, top 1%	93.57	94.50
Memory + reconstruction, maximum	99.11	99.21
Memory + reconstruction, top 0.1% (primary)	99.13	99.21
Memory + reconstruction, top 1%	99.11	99.17

Table 27: **Individual-seed scale and pooling controls on VisA.** Mean image AUROC (%) over all 12 classes and 2,162 official test images. Each row uses one fixed final model. Native and high-resolution crops are 392 and 518 pixels; both preserve the same field of view. Normal-calibrated readouts receive equal weight. These controls precede the two-model composition.

Seed	Native top 1%	High res. top 1%	Two scales top 1%	Two scales top 0.1%	Two scales two areas
1	99.09	99.17	99.31	99.15	99.32
2	99.39	99.28	99.48	99.34	99.51

Table 28: **Local memory composition on MVTec 3D-AD.** Mean image AUROC (%) over 10 classes and 1,197 test images. All readouts share the same fit/calibration images and memory banks. RGB+depth is the fixed equal mean of their separately calibrated evidence.

Readout	RGB	Depth	RGB+depth
Maximum	86.28	86.72	92.37
Top 0.1% (primary)	86.27	86.82	92.38
Top 1%	85.99	87.33	92.39

1% readout for 3D, and the normal-selected B/518 temporal stream for Ped2. The latter reaches 98.57 frame AUROC before any cross-stream cohort-rank fusion. Thus the overview documents benchmark coverage under stated settings; the paired experiments below measure changes within a fixed pipeline.

Table 29: **Cross-domain detector composition.** Each reference is the frozen predecessor in the corresponding experiment. Values are AUROC/AP (%); Δ is the paired difference in percentage points. FS+BT uses unlabelled evaluation-cohort ranks. Full per-domain comparisons appear in Appendix B.10.

Dataset	Prot.	ANP	Reference	Δ
MVTec-AD	FS	**99.54**/**99.86**	98.99/99.62	+0.55/+0.23
MVTec-AD (new)	FS	**99.88**/**99.97**	99.87/99.97	+0.01/+0.01
VisA	FS	**99.47**/**99.57**	99.41/99.53	+0.06/+0.04
BMAD	FS	**87.34**/**86.94**	87.31/86.57	+0.03/+0.37
MVTec 3D-AD	FS+BT	**97.45**/**99.35**	97.19/99.29	+0.26/+0.06
Ped2	FS+BT	**99.36**/**99.86**	99.34/**99.86**	+0.02/+0.00

The new MVTec-AD configuration anneals the reconstruction expert and retains continuous normal-fitted evidence; its reference is the three-role configuration before annealing. The 3D configuration uses equally weighted native and high-resolution reconstruction evidence; its reference is the native-resolution complete geometry composition. Both retain the same equal-rank geometry fusion. The VisA pair uses the complete configurations in Table 20.

The compiler logic is unchanged across MVTec-AD, VisA, and BMAD; only the available observation and expert primitives differ. On MVTec-AD, DINOv2-B reaches 98.95% mean image AUROC over 15 classes. A preregistered scale replication changes only the backbone to DINOv2-L and reaches 98.99%. The larger model improves ranking slightly but not its operating point: median test FPR rises from 3.10% to 5.00%, and the number of classes above $\alpha = 0.05$ grows from 5 to 7. Grid and wood show the same exchangeability pattern as pushpins, with quiet train-fold audits ($\leq 9.50\%$) but test FPR up to 42.00%. Because MVTec-AD has no validation split, these classes would need a separate deployment-normal audit.

In our first-party runs, DINOv2-L is the strongest strict single primitive; RD4AD, PatchCore, and EfficientAD reach 98.21%, 97.87%, and 94.89%. The three-role compiler converts DINOv2-L global structure, PatchCore local memory, and RD4AD reconstruction to normal-calibrated p-values and averages their log evidence. With fixed equal weights and no cohort normalization, it reaches 99.54% AUROC and 99.86% AP, improving DINOv2-L by 0.55 percentage points (paired-bootstrap 95% CI $[+0.05, +1.06]$ points). A four-role EfficientAD control reaches $99.47\%/99.84\%$, and the earlier DINOv2-L+PatchCore BT route reaches $99.42\%/99.81\%$. A generated geometric program does not transfer to this benchmark and lowers AUROC to 91.91%, serving as a fixed negative control in Table 30.

In the legacy VisA diagnostic, the structural primitive reaches 96.10% mean image AUROC and keeps the conformal flag rate at or below 0.06 on 11 of 12 classes. Cashew, chewinggum, and capsules reach 98.60%, 99.70%, and 97.30%, respectively; capsules falls back to this branch after synthesis fails. Published few-shot, training-free methods report 87–94 (Jeong et al., 2023; Gu et al., 2025), and full-shot trained detectors reach 99.3 (Guo et al., 2025), under their own protocols. The same strict three-role compiler used on MVTec-AD reaches 98.73% AUROC and 98.93% AP, a 2.65-point gain over the structural primitive (paired-bootstrap 95% CI $[+2.00, +3.35]$). The four-role control is slightly higher at $98.92\%/99.13\%$, but the three-role definition is kept fixed across both datasets.

Generated programs on VisA provide interpretable measurements but are not automatically useful fusion partners. Min-p fusion changes cashew from 98.60% to 97.10%, pcb1 from 98.60% to 96.60%, and pipe_fryum from 99.60% to 99.70%. The normal-only audit therefore retains the structural branch for all three, raising their subset mean from 97.80% under min-p fusion to 98.90%. The cashew program alone is stable across five seeds at 92.20%–94.00% AUROC with FPR at or below α, but it does not improve the stronger branch.

BMAD is a cross-domain test of complementarity after a strong detector has already been assembled. For each of BrainMRI, LiverCT, and RESC, the agent sees 12 induction normals and three image prompts, writes three `extract(gray)` candidates, and receives execution feedback on 30 disjoint verification normals. NumPy, OpenCV, and math are the only sandboxed libraries. Six of nine candidates execute successfully and five pass CFv3 stress execution. A five-seed, family-macro card selects one program per domain before the test manifest is opened. Separately, the compiler chooses a DINOv2-G/AnomalyMoE/PatchCore base from eight predeclared recipes. A program is admitted only with CFv3 gain at least 0.10 percentage points, nonnegative median gain, and nonnegative gain in at least four of five seeds. The selector then takes the smallest accepted intervention, a one-sided top-5% program tail at equal weight. The system uses no evaluation-cohort ranks and remains independent-image FS.

Adding the selected program improves AUROC and AP in each of the three domains. The macro AUROC change is small, $+0.03$ points, but is resolved by a paired within-class/label bootstrap (95% CI $[+3.00 \times 10^{-3}, +0.05]$). Mean AP improves by 0.37 points (CI $[+0.18, +0.57]$). We therefore treat BMAD as evidence that generated measurements add consistent tail information to an already strong three-expert compiler, rather than as the main effect-size result. The larger causal program effect is the LOCO removal result in Table 2. For calibrated BMAD decisions, the deployment audit is still required: with patient-disjoint train and test normals, it repairs LiverCT and RESC FPR from $0.29/0.27$ to $0.05/0.05$ using a separate normal batch.

B.11 Video and RGB–3-D Extensions

The compiler also accepts observations that are not static RGB images. For 3-D data, programs can read RGB features and point-cloud geometry; for video, they can operate on temporal score streams.

Table 30: First-party MVTec-AD executions on the same manifest. Values are mean-class image AUROC/AP. "Calibrated" denotes fixed equal log evidence from normal-only p-values; BT denotes percentile ranks over the unlabeled evaluation cohort. No test label enters a score or weight.

Method	Prot.	AUROC	AP
EfficientAD	FS	94.89	97.14
PatchCore	FS	97.87	99.33
RD4AD	FS	98.21	99.29
ANP DINOv2-L structural primitive	TF	98.99	99.62
ANP DINOv2-L + generated program	TF+BT	91.91	97.02
ANP DINOv2-L + PatchCore	FS+BT	99.42	99.81
ANP calibrated three-role compiler	FS	**99.54**	**99.86**

Table 31: BMAD image AUROC and mean image AP. Published rows are the five-run BMAD benchmark values (Bao et al., 2024); rerun rows use the same 7,013-image manifest. The full system is FS because its frozen base contains trained AnomalyMoE/PatchCore experts; the generated programs themselves use only normal images. Bold is best in each reported column.

Method	Source	Prot.	BrainMRI	LiverCT	RESC	Mean	Mean AP
SimpleNet	Pub.	FS	82.52	72.28	76.15	76.98	—
PatchCore	Pub.	FS	91.65	60.28	91.55	81.16	—
PatchCore	Ours rerun	FS	90.07	66.39	93.15	83.21	81.26
AnomalyMoE	Ours rerun	FS	92.40	71.64	92.47	85.50	84.84
ANP strong expert compiler	Ours	FS	92.60	74.55	94.79	87.31	86.57
ANP + agent-generated programs	Ours	FS	**92.64**	**74.56**	**94.81**	**87.34**	**86.94**

Normal references still define the programs, inference makes no language-model calls, and labels are used for evaluation. MVTec 3D-AD (Bergmann et al., 2022b) reports mean-class image AUROC/AP over ten categories, whereas Ped2 (Mahadevan et al., 2010) reports global frame AUROC/AP over 2,010 frames. These blocks test the breadth of the compiler contract and are not directly comparable to one another.

The first RGB+XYZ patch program on MVTec 3D-AD reaches 56.32% mean-class AUROC. Replacing its RGB representation with last-layer DINOv2-B features at 518 pixels and 160 normal references raises AUROC to 91.88%. Layers $\{5, 8, 11\}$ are available from the same forward pass; their predeclared top-1% readout gives a strict independent-image result of $93.27\%/98.19\%$ AUROC/AP. The batch-transductive route averages four pooling ranks (top 0.1%, 0.25%, 0.5%, and 1%) with one AnomalyMoE rank. It adds equal-weight geometry only when the training-normal alert prevalence satisfies $q_c = \Pr[p_{\text{geom}}(x) \le \alpha] \ge 4\alpha$. With all 210–361 normal references per class, the gate activates bagel, cookie, and rope and reaches $94.91\%/98.67\%$.

A five-seed CFv3 card independently selects a potato program using layers 5/8/11 and a top-0.25% readout. Adding this card to the all-reference base gives $95.07\%/98.71\%$, with no 3-D test label used to choose the class, readout, or weight. A second, uniform ten-class selector requires more than one point of mean synthetic gain and positive gain in four of five seeds. It preserves the existing geometry route and admits potato and tire to a pure FPFH program, using 0.5 base rank $+0.5$ FPFH rank for both. This rule reaches $95.85\%/98.91\%$, a 0.79-point mean AUROC gain over the all-reference route (paired-bootstrap 95% CI $[-0.02, +1.63]$ points). The point estimate is 0.70 points above CPMF's published 95.15% mean-class AUROC (Cao et al., 2024). The strict RGB map also reaches $98.73\%/50.05\%/95.25\%$ pixel AUROC/AP/AUPRO@0.3, 2.32 AUPRO points above CPMF's reported 92.93%.

The classwise replay explains where the gain comes from. Pure FPFH, with no rendered views, reaches $94.57\%/98.79\%$ on potato but only $75.08\%/90.73\%$ on tire. Equal fusion gives potato $94.54\%/98.64\%$ and lowers tire from $92.30\%/97.89\%$ to $90.46\%/97.08\%$. Tire remains in the automatic result because its eligibility and weight were fixed before evaluation. Replacing potato alone would yield the diagnostic $96.04\%/99.01\%$ row, but that is not the ten-class selector output. Cached $(784, 33)$ FPFH probes were checked elementwise against CPMF's native preprocessing before full scoring.

On Ped2, our AnomalyMoE replay reaches $97.19\%/99.41\%$ frame AUROC/AP and the DINOv2-B diffuse-map stream reaches $98.45\%/99.65\%$. Five-seed normal-only selection chooses `map-mean + Gaussian-5`. The L/518 branch also keeps an independently selected `top-20% + Gaussian-5` readout and averages the two program ranks. Equal weighting of the frozen B/518, L/518, and G/518 branches, followed by one MoE rank, reaches $99.34\%/99.86\%$.

The same CFv3 receipts provide a precision estimate for each program. Inverse synthetic-AUROC variance gives B/L/G weights of approximately $0.29/0.33/0.38$; the subsequent 0.5 DINO–0.5 MoE composition reaches $99.36\%/99.86\%$. Programs and weights do not read Ped2 labels, although final percentile ranks use the unlabeled cohort, hence FS+BT. A 10,000-replicate sequence bootstrap places the precision-over-equal AUROC change at $+0.02$ percentage points with CI $[-0.04, +0.08]$ points. The weighted point is the frozen output, but the data do not distinguish it statistically from equal weighting.

Table 32: Video and RGB–3-D extensions. MVTec 3D-AD values are mean-class image AUROC/AP; Ped2 values are global frame AUROC/AP. BT denotes within-class unlabeled evaluation-cohort ranks; Pub. denotes a published comparator rather than our rerun.

Dataset	Compiled route	Prot.	AUROC	AP
MVTec 3D-AD	RGB+XYZ patch program	FS	56.32	83.40
MVTec 3D-AD	DINOv2-B 518, 160 refs	FS	91.88	97.73
MVTec 3D-AD	$\{5, 8, 11\}$ strict top-1%	FS	93.27	98.19
MVTec 3D-AD	ten-class CFv3 route, $k = 160$	FS+BT	94.63	98.61
MVTec 3D-AD	all-ref. + frozen potato program	FS+BT	95.07	98.71
MVTec 3D-AD	CPMF (Cao et al., 2024)	Pub.	95.15	—
MVTec 3D-AD	CFv3-gated equal FPFH compiler	FS+BT	**95.85**	**98.91**
Ped2	AnomalyRuler (Yang et al., 2024)	Pub.	97.90	—
Ped2	AnomalyMoE frame score	FS	97.19	99.41
Ped2	DINOv2-B diffuse-map score	FS	98.45	99.65
Ped2	Gaussian-1 equal baseline	FS	98.91	99.76
Ped2	typed cross-stream program (DINO-B)	FS	99.07	99.79
Ped2	typed cross-stream program (DINO-L)	FS	99.09	99.80
Ped2	fixed B/L cross-backbone rank fusion	FS	99.10	99.80
Ped2	fixed B/L/G modality-balanced fusion	FS+BT	99.34	99.86
Ped2	CFv3-precision B/L/G compiler	FS+BT	**99.36**	**99.86**

Table 33: First-party cross-modal audit on the two classes admitted by the frozen CFv3 eligibility gate. The potato-only row is an explicitly labeled diagnostic and is not the automatic ten-class result.

Route	Potato AUROC/AP	Tire AUROC/AP	Ten-class AUROC/AP
ANP all-reference route	84.83/95.85	92.30/97.89	95.07/98.71
CPMF pure FPFH	94.57/98.79	75.08/90.73	—
Frozen equal fusion	94.54/98.64	90.46/97.08	**95.85/98.91**
Potato-only replacement (diagnostic)	94.57/98.79	—	96.04/99.01

Table 34: MVTec 3D-AD pixel localization. Our RGB map reuses the strict DINO forward; CPMF is the repository-reported comparator.

Method	Pixel AUROC	Pixel AP	AUPRO@0.3
CPMF (Cao et al., 2024)	—	—	92.93
ANP strict RGB map	**98.73**	**50.05**	**95.25**

B.11.1 A stronger reconstruction expert

We also train a Dinomaly2-derived reconstruction expert for 40,000 updates on normal MVTec-AD images, using a held-out normal split for calibration and the final checkpoint for evaluation.

Native top-1% pooling reaches 99.87% mean AUROC. A normal-fitted Yeo–Johnson transformation followed by Gaussian log-tail evidence for the two reconstruction layers reaches 99.87%. The empirical-tail controls are reported alongside these continuous scores in Table 35.

Table 35: **Reconstruction-expert study on MVTec-AD.** Mean image AUROC (%) over all 15 classes and 1,725 test images. The new expert uses the final checkpoint after 40,000 normal-only training updates. All readouts and the fixed role-composition controls are retained.

Configuration	AUROC
Historical three-role configuration	99.54
New reconstruction, native	99.87
New reconstruction, continuous layer calibration	99.87
New reconstruction, empirical layer calibration	96.76
New reconstruction, empirical native calibration	94.69
Replace reconstruction role	99.77
Add reconstruction role	99.71
Continuous three-role composition	99.87
Continuous two-role control	99.80

A fixed three-role composition replaces the historical RD4AD evidence with the new expert's empirical log-tail evidence, retaining DINO global structure and PatchCore local memory. It improves the historical 99.54% to 99.77% (paired gain $+0.23$ points, 95% CI $[+0.07, +0.45]$). Adding the new expert as a fourth role gives 99.71%. The results use one completed detector seed and inherited historical expert assets; they are a development study rather than an exact reproduction of the published Dinomaly2 setting. An independent audit reconstructs both calibration readouts, checks every test identity and replays 8,000 paired bootstrap values. These expert results are separate from evidence about agent-written program generation.

To preserve ordering beyond the largest calibration observation, we fit Yeo–Johnson standardization to the DINO and new reconstruction scores on their respective normal references and use $\log(1 + \exp(z))$ as continuous evidence. The fixed equal three-role mean retains PatchCore's original evidence and reaches 99.87%, improving the historical configuration by 0.33 points (95% CI $[+0.13, +0.64]$). Its difference from the new native expert is $+4.00 \times 10^{-3}$ points (95% CI $[-0.11, +0.13]$). The two-role DINO–reconstruction control reaches 99.80%. This follow-up was motivated by observed tail saturation in the development results. Independent score reconstruction and paired resampling reproduce all reported values; no normal-training image hash overlaps a test image.

We repeat reconstruction training with a second initialization and retain the same class-normal fit and calibration split. Table 36 reports both complete runs and their predeclared equal ensemble. The ensemble reaches 99.89% with reconstruction alone and 99.87% with the fixed three-role composition. Relative to the historical three-role configuration, the latter gains 0.34 points (paired 95% CI $[+0.14, +0.64]$). Its difference from reconstruction alone is -0.01 points (95% CI $[-0.12, +0.11]$). These image-resampling intervals condition on the two trained models. The main overview retains the single-model configuration.

Table 36: **Fixed reconstruction repeats on MVTec-AD.** Mean image AUROC (%) over all 15 classes. Both models use 40,000 updates and identical fit/calibration identities. The ensemble averages separately normal-calibrated reconstruction evidence; the inherited appearance experts remain fixed.

Configuration	Seed 1	Seed 2	Equal ensemble
Reconstruction	99.87	99.87	99.89
Three-role composition	99.87	99.87	99.87
Two-role control	99.80	99.75	99.77

We further test a fixed cosine-decay tail, reducing the learning rate to 1% of its original endpoint over 10,000 additional updates. The same 3,269 normal training images and 360 calibration images are retained. Replacing only the reconstruction expert yields 99.88% AUROC for the three-role

composition, with a paired change of $+0.01$ points from its predecessor (95% CI $[-0.02, +0.04]$). Table 37 retains the complete expert and composition controls. The model is evaluated at the fixed final update.

Table 37: **Fixed annealing on MVTec-AD.** Mean image AUROC and AP (%) over all 15 classes. The final model receives 10,000 cosine-decay updates after the original 40,000 updates; the normal fit/calibration split remains fixed.

Configuration	AUROC	AP
Original reconstruction	99.87	99.96
Annealed reconstruction	99.90	99.98
Annealed, normal calibration	99.90	99.98
Annealed, layer calibration	99.89	99.97
Original + annealed, equal evidence	99.90	99.98
Annealed, continuous evidence	99.90	99.98
Original three-role composition	99.87	99.97
Annealed three-role composition (primary)	99.88	99.97

The reconstruction experiment also covers MVTec 3D-AD. We train on 2,394 normal RGB–depth pairs and retain 262 normal pairs for calibration. The RGB and HOT-coloured depth images share plane removal and gap filling; the encoder fuses their features with equal weight. Evaluation uses the final 40,000-update model on all 1,197 test images. Native top-1% pooling reaches 96.23% AUROC and 99.00% AP. Table 38 reports the fixed layer-calibration controls. These results evaluate the reconstruction expert and its readout, while the earlier RGB and RGB+3D compositions retain their own configurations. The normal calibration split and final-checkpoint rule remain distinct from the published reference setting.

Table 38: **Reconstruction readouts on MVTec 3D-AD.** Mean image AUROC and AP (%) over all 10 classes and 1,197 test images. All rows use the same final model and RGB+depth observations.

Readout	AUROC	AP
Native top-1%	96.24	98.99
Continuous layer calibration	96.15	98.98
Empirical layer calibration	95.76	98.66

Combining the reconstruction expert with the local RGB and depth memories gives 95.72% AUROC, a gain of 3.34 points over the two-memory control (paired 95% CI $[+2.47, +4.24]$). The difference from reconstruction alone is -0.51 points (95% CI $[-1.25, +0.17]$). All roles receive equal weight; Table 39 retains every predeclared memory readout. The stronger reconstruction expert therefore provides most of the detection performance in this configuration. These intervals resample images within class and label, conditional on the fitted experts.

Table 39: **Fixed appearance–reconstruction fusion on MVTec 3D-AD.** Mean image AUROC (%) over all 10 classes. Each role uses normal-fitted log calibration, and role evidence is averaged with equal weights.

Memory readout	RGB + reconstruction	RGB + depth + reconstruction
Maximum	95.47	95.71
Top 0.1% (primary)	95.51	95.72
Top 1%	95.51	95.75

We examine scale sensitivity by averaging evidence from the full image and four overlapping 392×392 tiles of a 588×588 image. RGB and depth share crop coordinates, overlapping anomaly maps are averaged, and each view is calibrated on the same held-out normal pairs. The fixed equal composition reaches 96.19% AUROC (Table 40); its difference from the native view is -0.05 points, with paired 95% CI $[-0.81, +0.71]$. Thus, the additional views preserve the aggregate performance within sampling uncertainty, while the single-image readout remains the more economical choice.

Table 40: **Full-image and overlapping-tile reconstruction on MVTec 3D-AD.** Mean image AUROC and AP (%) over all 10 classes, using the same final model. The native and full-image views coincide at the 3D model's 392×392 input size.

View composition	AUROC	AP
Native	96.24	98.99
Full image	96.24	98.99
Overlapping tiles	90.27	97.04
Full image + tiles (primary)	96.18	99.01

The historical geometry composition and the new reconstruction expert provide complementary evidence. We retain the historical CFv3 gate and combine its complete output with native reconstruction using the same equal-rank rule for every class. This FS+BT composition reaches 97.19% AUROC and 99.29% AP (Table 41), improving the historical composition by 1.34 points (paired 95% CI $[+0.62, +2.10]$) and reconstruction alone by 0.96 points (95% CI $[+0.25, +1.68]$). The two-scale extension below is the configuration in Table 1. Its ranks depend on the unlabelled evaluation cohort, and its expert assets retain their original normal-reference allocations. The intervals condition on those fitted assets. The result establishes complementarity in this development setting; the separate program-generation experiments address the contribution of automatic synthesis.

Table 41: **Complementary geometry and reconstruction on MVTec 3D-AD.** Mean image AUROC and AP (%) over all 10 classes. The new composition averages the two within-class percentile ranks with fixed equal weights.

Configuration	Protocol	AUROC	AP
Historical geometry composition	FS+BT	95.85	98.91
Reconstruction expert	FS	96.24	98.99
Geometry + reconstruction	FS+BT	**97.19**	**99.29**

At 518-pixel resolution, we also estimate the normal reconstruction mean and variance at each image location using fitting-normal images. A fixed 10% variance shrinkage toward the spatial average regularizes the standardized residual. Independent normal calibration then combines uncorrected and residual evidence equally. The complete controls appear in Table 42.

Table 42: **High-resolution spatial residual calibration on MVTec 3D-AD.** Mean image AUROC and AP (%) across all classes at 518-pixel input resolution. Spatial means and regularized variances use only training-normal images. Native and standardized residual evidence receive equal weight, calibrated on disjoint normal images. Geometry compositions use FS+BT.

Configuration	AUROC	AP
Current two-scale complete composition	97.45	99.35
Uncorrected high-resolution reconstruction	96.85	99.16
Spatially standardized reconstruction	94.92	98.31
High-resolution + spatial reconstruction	96.68	99.00
Geometry + high resolution	97.48	99.35
Geometry + spatial residual	96.83	99.18
Geometry + high/spatial (primary)	97.37	99.32

We also estimate the normal reconstruction mean and variance at each image location using fitting-normal images. A fixed 10% variance shrinkage toward the spatial average regularizes the standardized residual. Independent normal calibration then combines native and residual evidence equally. The complete controls appear in Table 43.

We increase the reconstruction input from 392 to 518 pixels while preserving the resize-to-crop ratio and orientation. Each scale uses the same held-out normal calibration, and their evidence receives equal weight. The complete composition and its fixed controls appear in Table 44.

Table 43: **Normal spatial residual calibration on MVTec 3D-AD.** Mean image AUROC and AP (%) across all classes. Spatial means and regularized variances use only training-normal images. Native and standardized residual evidence receive equal weight, calibrated on disjoint normal images. Geometry compositions use FS+BT.

Configuration	AUROC	AP
Native reconstruction	96.24	98.99
Spatially standardized reconstruction	96.58	98.95
Native + spatial reconstruction	96.62	99.07
Geometry + native	97.19	99.29
Geometry + spatial residual	97.39	99.32
Geometry + native/spatial (primary)	97.37	99.33

Table 44: **Full-context two-scale inference on MVTec 3D-AD.** Mean image AUROC and AP (%) across all classes. The same final model uses 392- and 518-pixel crops with a fixed field of view; the two calibrated scores receive equal weight. Geometry compositions use FS+BT.

Configuration	AUROC	AP
Native reconstruction	96.24	98.99
High-resolution reconstruction	96.85	99.16
Two-scale reconstruction	96.89	99.17
Geometry + native	97.19	99.29
Geometry + high resolution	97.48	99.35
Geometry + two scales (primary)	97.45	99.35

The additional 10,000 reconstruction updates use all 2,656 official training-normal pairs, including the former calibration subset; this variant has no independent calibration claim. We retain the same equal-rank geometry composition, which reaches 97.15% AUROC; its difference from the original complete composition is -0.04 points (paired 95% CI $[-0.26, +0.17]$). The fixed final checkpoint and all complete composition controls appear in Table 45.

Table 45: **Geometry composition after fixed reconstruction annealing.** Full normal-pool refit. Mean image AUROC and AP (%) over all 10 MVTec 3D-AD classes. Geometry compositions use the same FS+BT rule and retain the historical geometry gate.

Configuration	AUROC	AP
Historical geometry	95.85	98.91
Annealed reconstruction	96.13	98.97
Geometry + original reconstruction	97.19	99.29
Geometry + annealed reconstruction (primary)	97.15	99.28
Geometry + two reconstruction models	97.15	99.28

The additional 10,000 reconstruction updates retain 2,394 normal fitting images and 262 independent normal calibration images. We retain the same equal-rank geometry composition, which reaches 97.06% AUROC; its difference from the original complete composition is -0.13 points (paired 95% CI $[-0.35, +0.06]$). The fixed final checkpoint and all complete composition controls appear in Table 46.

We also average predictions over rotations and reflections at the native image scale. The eight anomaly maps are spatially realigned before averaging and top-1% pooling. Each readout uses the same held-out normal calibration; the final geometry composition retains equal within-class ranks. It reaches 92.36% AUROC, a change of -4.84 points from the native-view geometry composition (paired 95% CI $[-6.13, -3.66]$). Table 47 retains rotation-only and score-averaging controls. The eight-view reconstruction requires additional inference, and its timing excludes the geometry branch.

Table 46: **Geometry composition after fixed reconstruction annealing.** Retained normal calibration. Mean image AUROC and AP (%) over all 10 MVTec 3D-AD classes. Geometry compositions use the same FS+BT rule and retain the historical geometry gate.

Configuration	AUROC	AP
Historical geometry	95.85	98.91
Annealed reconstruction	95.98	98.93
Geometry + original reconstruction	97.19	99.29
Geometry + annealed reconstruction (primary)	97.06	99.25
Geometry + two reconstruction models	97.13	99.27

Table 47: **Orientation averaging on MVTec 3D-AD.** Mean image AUROC and AP (%) over all 10 classes. The same final reconstruction model processes the native image, four rotations, or all eight rotations and reflections.

Readout or composition	AUROC	AP
Native reconstruction	96.24	98.99
Rotation-averaged maps	82.73	94.35
Dihedral-averaged maps	82.12	93.99
Rotation-averaged scores	85.63	95.60
Dihedral-averaged scores	84.64	95.26
Geometry + native reconstruction	97.19	99.29
Geometry + rotation maps	92.64	97.89
Geometry + dihedral maps (primary)	92.36	97.77

B.12 Efficiency

Table 48: First-party latency with an RTX A6000, LOCO breakfast_box, batch 1, and CUDA synchronization. Core is model/mask-ready inference; online adds image decoding, preprocessing, and transfer. ANP and AnomalyMoE use 5 warm-up + 80 measured images; cached-SAM uses 5 + 20. Model/reference setup and pre-generated score-free masks are excluded where applicable. The ANP row is the configuration reported for accuracy in Table 2.

Method	Online LLM	Core mean (ms)	Online mean (ms)	n
ANP (the configuration reported for accuracy)[a]	**0**	**104.10**	**191.13**	80
AnomalyMoE (Gu et al., 2026)	0	130.81	334.04	80
LogSAD (Zhang et al., 2025a) (cached SAM)[b]	0	1360.21	1677.36	20
UniVAD (Gu et al., 2025)	0	8879.98	9044.88	10

[a] core includes CLIP/DINOv2, score-free mask consumption, instance/structural scoring, and fusion.
[b] scores match direct exactly (max. $|\Delta| = 0$, atol 10^{-12}); direct: 4379.11/4664.27 ms.
Only SAM generation is cached; decoding, features, region processing, and fusion remain online.

On the 80-image run, the same ANP point used for the accuracy table is $1.26\times$ faster than AnomalyMoE in model-ready computation and $1.75\times$ faster from decoding to score. Against cached-SAM LogSAD, the factors are $13.1\times$ and $8.78\times$. Its mean stage times are 86.63 ms for decoding/preprocessing, 0.41 ms for host-to-device transfer, and 104.10 ms for the synchronized core; all 85 warm-up and measured lookups hit the pre-generated component cache.

The cache boundary follows the same convention used by component-mask methods: mask generation and model/reference setup are offline, while consuming masks, extracting CLIP/DINOv2 features, matching instances, scoring structure, and fusion remain online. The cached masks contain no scores, features, labels, or decisions. LogSAD's older cached-SAM row is also retained, including its direct score-equivalence receipt, so the gain is not created by silently dropping a scoring stage.

Two faster numbers are deliberately kept out of the primary comparison. The lightweight ANP-fast ablation takes $29.55/119.30$ ms core/online but reaches only 81.50 mean AUROC. The EfficientAD juice component takes $5.26/54.90$ ms, but it is one expert rather than the five-class detector. Neither is paired with the 94.34 accuracy result. AnomalyClaw and AgentIAD are omitted because we have

not executed them under this timing protocol; their published call counts are used only as context in Section 1.

B.13 Visual Comparison of Frozen Configurations

Figures 8 and 9 use every official test image and the same class manifests for all configurations. They resolve the aggregate ablation by class and show its effect across operating points. Figure 10 pairs full-resolution benchmark images with the corresponding frozen scores. Each displayed score is its percentile among that configuration's test-normal scores, with half credit for ties. These retrospective percentiles are not used to fit calibration or to make deployed decisions. The first two logical-anomaly IDs in each category are shown. Original test photographs are matched by SHA256 to the archived evaluation inputs; defect annotations are used only for display.

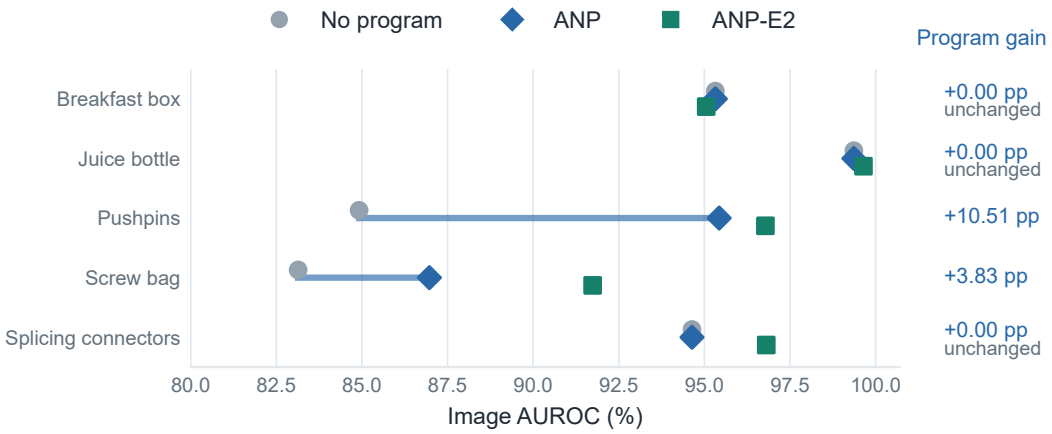

Figure 8: **Selective gains without observed classwise AUROC regression.** Markers show pooled image AUROC for the complete five-class test set. Connecting segments and right-side labels give ANP minus the matched no-program pipeline. The three inactive categories preserve every detector score (+0.00 pp); pushpins and screw bag gain +10.51 and +3.83 pp. The mean gain is 2.87 percentage points. ANP-E2 is a separate ensemble configuration, outside this program-layer ablation.

Figure 11 exposes the perceptual representations underlying two of the evaluated routes. The differences include how tray regions, bag background, and small components are partitioned. This comparison helps interpret why component-level checks depend on the measurement interface; it does not assign a segmentation-accuracy ranking to either model.

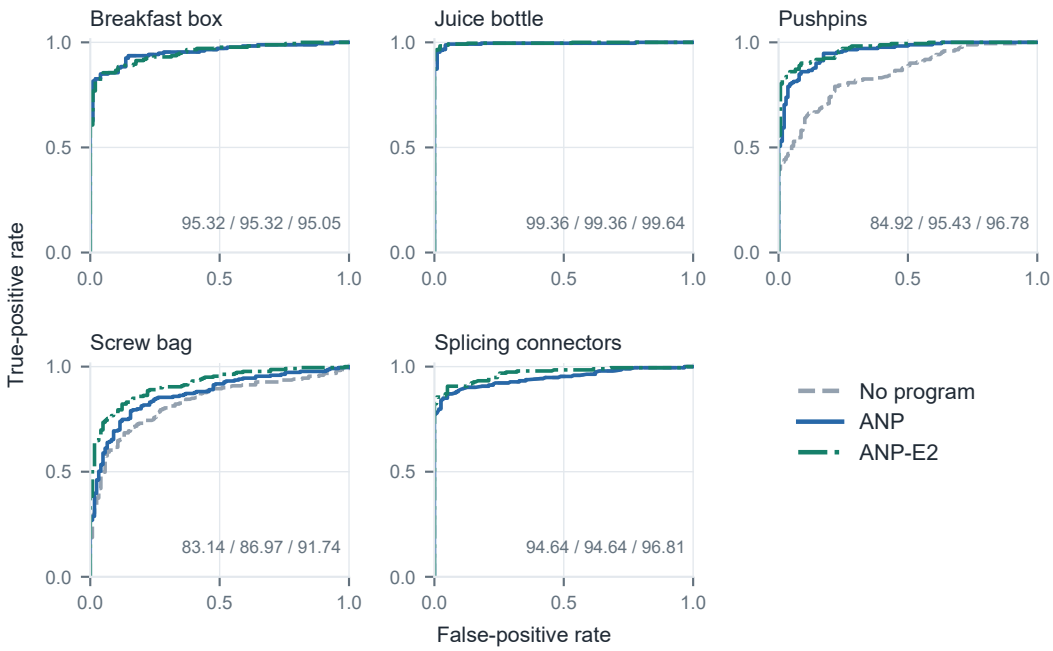

Figure 9: **Complete per-category LOCO ROC curves.** All 1,568 normal, logical-anomaly, and structural-anomaly test images are included. Colors and line styles identify the same configurations throughout. Panel annotations give AUROC (%) in the order no program / ANP / ANP-E2.

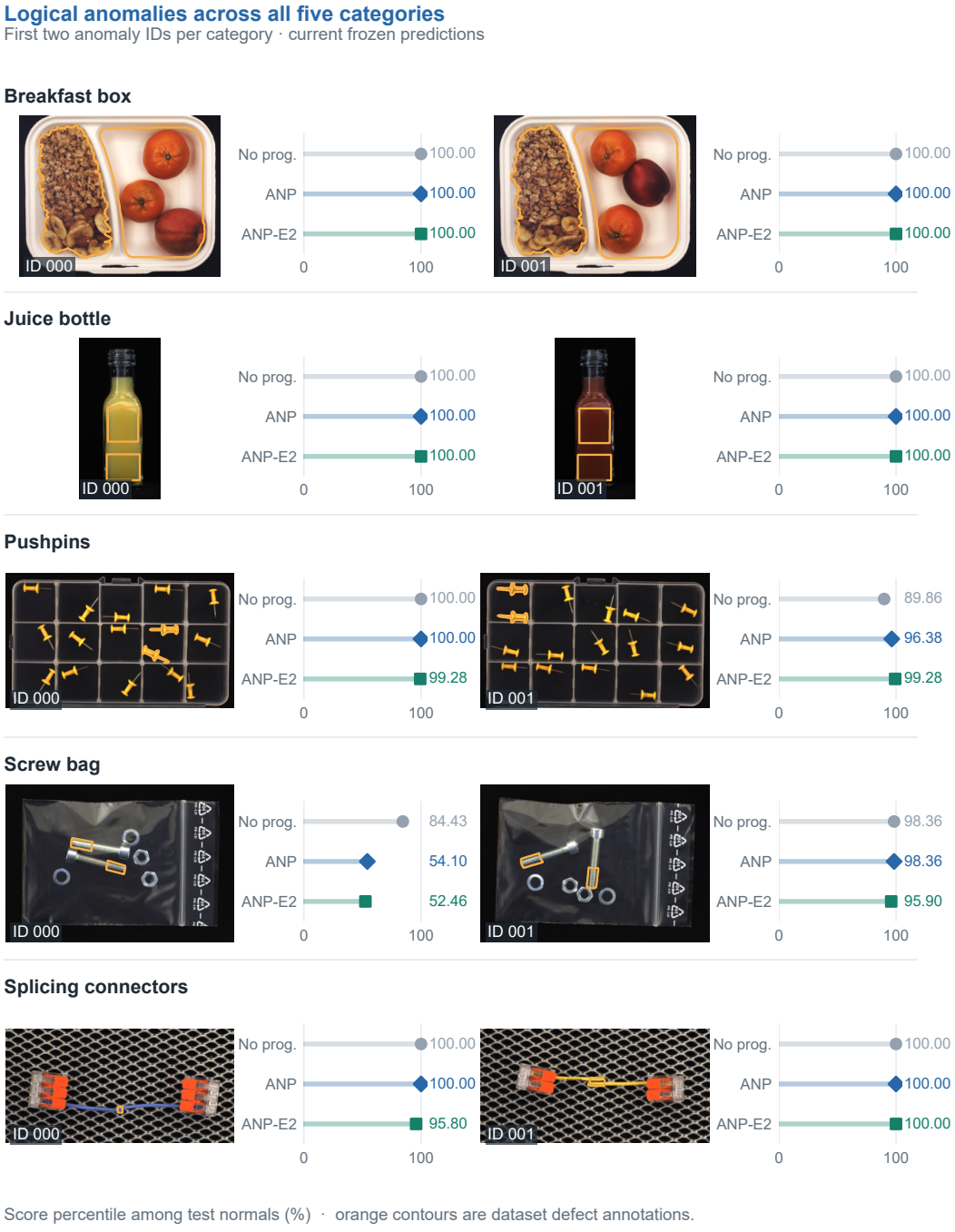

Figure 10: **Logical anomalies and their recorded detection scores.** Each category shows anomaly IDs 000 and 001 with dataset defect contours in orange. Adjacent markers give the test-normal score percentiles of the no-program pipeline, ANP, and ANP-E2 on identical images. Larger values indicate higher anomaly ranking. The fixed selection includes both strong rankings and the screw_bag examples discussed in the verification analysis.

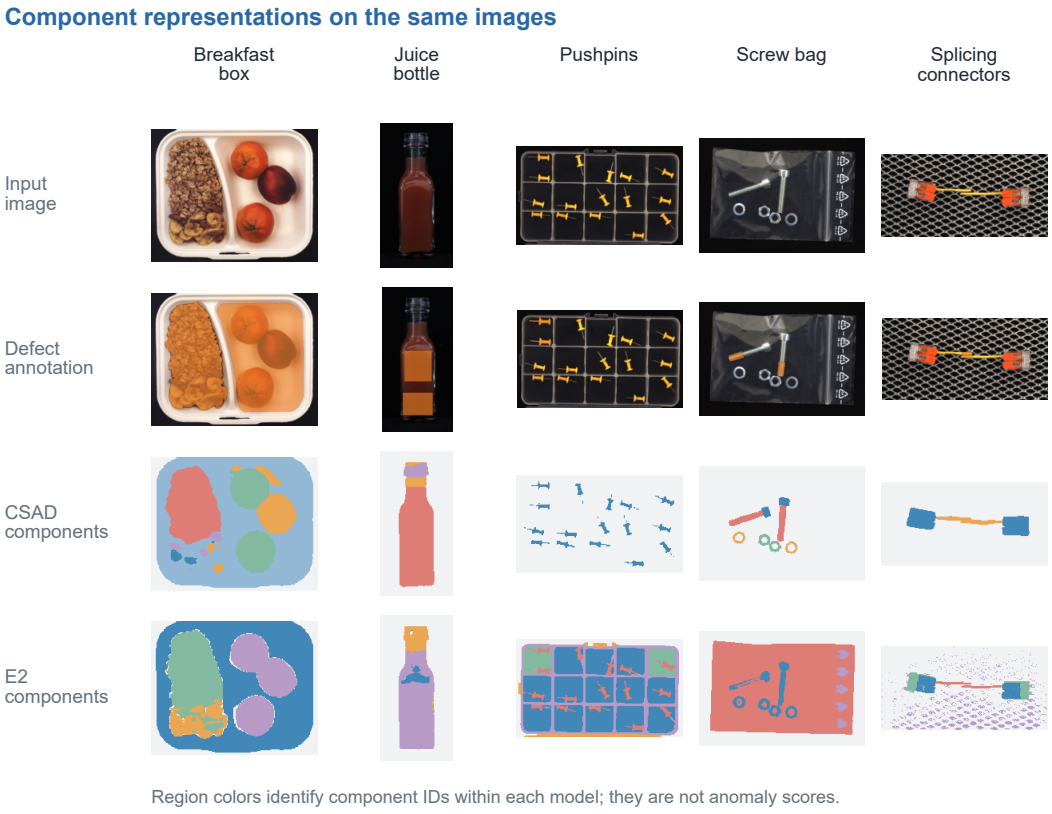

Figure 11: **Different component representations of identical inputs.** Columns show logical-anomaly ID 001 for each category. The second row overlays dataset defect annotations. The last two rows are archived component maps from our CSAD author-checkpoint execution and the SALAD-derived component path used by ANP-E2, respectively. Each map is checked against its recorded hash. Colors indicate component IDs within a model and do not imply a correspondence between models. These are perception outputs, not anomaly heatmaps; their downstream detection scores are shown in Figure 10.