# TRIAD: CALIBRATING CONTEXTUAL AND NORMAL-SUPPORT EVIDENCE FOR UNSUPERVISED MULTI-CLASS ANOMALY DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Unified multi-class anomaly detection asks a single model, trained only on normal images, to detect and localize defects across many product categories. Scoring pipelines that inspect the target they are scoring suffer from two structural failures. First, attention lets target content inform its own score, so a plausible defect can be explained as normal. Second, raw error maps lack a common probabilistic scale across positions, appearance modes, and categories. We present TRIAD (Target-blind energy, Reference witness, and Integrated Anomaly Detection), a framework built entirely on a frozen DINOv2 backbone that computes three calibrated evidences for every test image. A target-blind contextual energy fills in masked patches from context alone and grades the observation by its likelihood under a von Mises–Fisher mixture over a library of normal patch atoms. A witness branch measures the distance to a coverage-maximizing library of normal images, aggregates the distance with a CVaR tail readout, and converts the aggregate into route-conditional empirical $p$-values with finite-sample false-alarm control. A calibrated fusion places both fields on one commensurable scale and combines them with equal weights, a statistic that is likelihood-ratio optimal for dense shifts under a Gaussian working model, while image-level decisions remain identical to the witness detector. On MVTec-AD and VisA under the unified protocol, TRIAD advances the state of the art on all seven reported metrics, reaching $98.49/98.97$ pixel AUROC, $68.25/46.61$ pixel AP, and $96.61/94.57$ AUPRO.

## 1 INTRODUCTION

Industrial quality inspection increasingly calls for unified multi-class anomaly detection. A single model must handle many product categories at once, train only on normal images, and return both an image-level decision and a pixel-level localization for any category at test time. This setting removes the per-category specialist assumption and matches production lines where defect types and product mixes evolve continuously.

Reconstruction- and masked-modeling-based pipelines dominate the current state of the art (You et al., 2022; He et al., 2024b; Zhang et al., 2023b; He et al., 2024a; Lu et al., 2026). We argue that two of their failures are structural rather than matters of tuning. The first failure concerns what the scorer sees. In a reconstruction pipeline the network that grades a patch also observes that patch, and attention supplies the target's own content to its score. A defect that is locally plausible can therefore be reconstructed faithfully and reported as normal. The second failure concerns what the score means. Raw reconstruction errors carry position-, mode-, and category-dependent offsets, so a single error magnitude means different things at different places on different products. Comparisons across positions and thresholds across categories lose their meaning unless the scores are calibrated.

This paper presents TRIAD (Target-blind energy, Reference witness, and Integrated Anomaly Detection), a framework that addresses both failures directly and computes three cooperating evidences on a fully frozen DINOv2 backbone (Oquab et al., 2024). Figure 1 gives an overview. The first evidence is a target-blind contextual energy. A multi-scale masking schedule hides every token exactly once per scale, a shallow decoder fills the masked positions from context alone, and the clean observation is graded by its likelihood under a von Mises–Fisher mixture over a library of normal patch



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.

atoms. No pathway exists through which the target informs its own score. The second evidence is a witness distance. A coverage-maximizing library of normal images records what the normal world actually contains, and each whole-view patch is scored by its distance to the closest normal patch. A CVaR tail readout aggregates the distance field into an image score, which is converted into route-conditional empirical $p$-values with a finite-sample false-alarm guarantee. The third contribution is calibrated fusion. Route-level statistics and a validation-derived scale correction place the two spatial fields on one commensurable scale, after which equal weights are provably optimal for the dense-shift alternative we care about. The witness field, which the scalar readout alone would discard, becomes a second localization evidence, and image-level decisions remain bit-identical to the witness detector, so fusion can only improve localization and never perturbs detection.

Our contributions are as follows.

- **Target-blind contextual energy.** Reconstruction-style scorers grade a patch while observing it, so attention can explain a locally plausible defect as normal. We build the query for every position without any access to the content it scores, and we grade the clean observation by its likelihood under a context-conditioned von Mises–Fisher mixture over a library of normal patch atoms. Owner exclusion during training and blind encoding at inference remove target leakage by construction rather than by regularization.
- **Witness evidence with distribution-free calibration.** A contextual signal cannot tell whether an appearance occurs anywhere in the normal world. We therefore ground scoring in a coverage-maximizing library of normal images, measure a nearest-neighbour witness distance for every whole-view patch, aggregate the distance field with a CVaR tail readout, and convert the aggregate into route-conditional empirical $p$-values under the add-one rule. The image-level decision inherits a finite-sample false-alarm guarantee.
- **Calibrated joint evidence.** The first two evidences are heterogeneous spatial fields: they are not directly comparable in raw form, and the witness field is aggregated away at the image level. We make the two fields commensurable through route-level calibration with validation self-standardization and fuse them with equal weights, which we show is likelihood-ratio optimal for dense shifts under a Gaussian working model. The fusion recovers a second localization evidence at no detection risk, because image decisions remain identical to the witness detector.

## 2 Related Work

**Unified multi-class anomaly detection.** Early reconstruction systems such as RD4AD (Deng & Li, 2022) and DeSTSeg (Zhang et al., 2023c) train one model per category. UniAD (You et al., 2022) established the unified protocol, and DiAD (He et al., 2024b), ViTAD (Zhang et al., 2023b), MambaAD (He et al., 2024a), CNC (Wang et al., 2025), and MaskAD (Lu et al., 2026) pushed the frontier with stronger backbone training and masked-modeling objectives. All of these methods train their backbones and grade a reconstruction of the target; TRIAD keeps the backbone frozen, never scores a reconstruction, and adds calibrated evidences.

**Memory- and embedding-based detection.** PatchCore (Roth et al., 2022) and PaDiM (Defard et al., 2021) score deviations against a memory of normal patch features, and CLIP- or DINOv2-based methods such as WinCLIP (Jeong et al., 2023) and AnomalyDINO (Damm et al., 2025) exploit frozen foundation features. TRIAD differs in three ways: it scores contexts rather than targets, it covers the normal appearance space at the image level, and it attaches finite-sample guarantees to the image decision.

**Masked image modeling.** MAE (He et al., 2022) and DINO (Caron et al., 2021) established masked pre-training. Our schedule is a multi-scale checkerboard probe with a seeded random shift that masks every token exactly once per scale, so the energy field has no sampling bias across positions.

**Calibration and conformal testing.** The add-one empirical $p$-value (Vovk et al., 2005) and inductive conformal testing (Papadopoulos et al., 2002) provide distribution-free validity under exchangeability. We apply them route-conditionally to an image score and extend calibration to the spatial field itself.

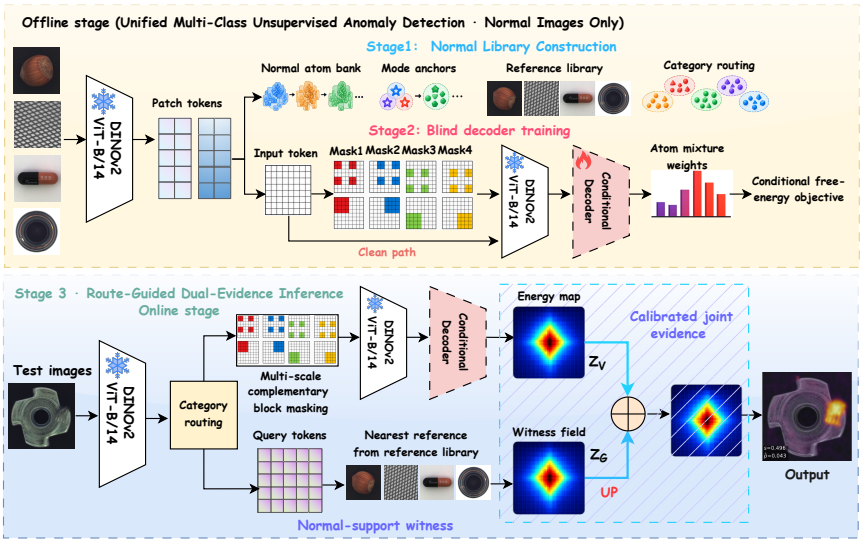


Figure 1: The TRIAD framework. Stage 1 constructs the per-category libraries on normal training images only: the frozen DINOv2 backbone provides patch tokens, from which the normal atom bank, the mode anchors, the reference library, and the category router are fitted. Stage 2 trains the blind decoder on normal images: masked-context queries are decoded into atom mixture weights, and the conditional free-energy objective optimizes the only trainable module. Stage 3 performs route-guided dual-evidence inference: category routing selects all per-route artifacts, the contextual energy field and the witness field are computed and calibrated, and their fusion yields the pixel map while the image decision remains identical to the witness branch.

# 3 METHOD

## 3.1 PROBLEM SETTING AND THE SHARED FROZEN ENCODER

A training set of normal images spans $C$ categories. One model must, at test time, return an image-level decision and a pixel-level map for any category. Every input image is bicubically resized onto an $896 \times 896$ canvas $x$, which is tiled into four native $448 \times 448$ views; the canvas doubles the linear inspection resolution of a single $448^2$ pass while keeping every encoder forward at the input size the backbone was trained on, and the mosaic lets each view serve as context for the other three. A frozen DINOv2 ViT-B/14 (Oquab et al., 2023; Dosovitskiy et al., 2021) extracts token features from layers $\mathcal{L} = \{6, 9, 12\}$ of every view. For each token $t$ we form a unit descriptor by $L_2$-normalizing every layer and concatenating,

$$z(t) \;=\; \tfrac{1}{\sqrt{L}}\big[\hat{h}_6(t);\, \hat{h}_9(t);\, \hat{h}_{12}(t)\big] \in \mathbb{R}^{2304}, \qquad \|z(t)\|_2 = 1, \tag{1}$$

so that inner products between descriptors average per-layer cosine similarities, where $\hat{h}_l$ denotes the $L_2$-normalized features of layer $l$ and $L = |\mathcal{L}| = 3$ is the number of concatenated layers. Routing uses a dedicated image-level semantic descriptor built from the last layer alone. The whole canvas is resampled to $448^2$ and passed through the backbone once; the mean of the resulting $T = 1024$ last-layer tokens is $L_2$-normalized,

$$s(x) \;=\; \frac{\frac{1}{T}\sum_{t=1}^{T} h_{12}(t)}{\left\|\frac{1}{T}\sum_{t=1}^{T} h_{12}(t)\right\|_2} \;\in\; \mathbb{R}^{768}, \tag{2}$$

where $h_{12}(t)$ is the last-layer token at position $t$. A single route per image is obtained by matching $s(x)$ against frozen class centroids $\nu_k \in \mathbb{R}^{768}$ estimated on the normal training split,

$$c(x) \;=\; \arg\max_{k}\; \langle s(x),\, \nu_k \rangle\,. \tag{3}$$

The route $c(x)$ indexes every per-category artifact used below: the atom bank, the mode anchors, the reference library, and the route-conditional nulls. Routing is deterministic and frozen.

### 3.2 TARGET-BLIND CONTEXTUAL ENERGY

**Why blindness matters.** In reconstruction-based scorers the network that grades a patch also sees it, and attention supplies the target's own content to its score. We remove this pathway by construction.

**Multi-scale masking schedule.** Each $448 \times 448$ view contributes $32 \times 32$ tokens at patch size $14$, so the four-view canvas yields a $64 \times 64$ token grid. On this grid, a schedule with block scale $s \in \mathcal{S} = \{2, 4\}$ partitions the canvas into blocks of $s \times s$ tokens and assigns the blocks to four mutually complementary mask classes arranged in a $2 \times 2$ checkerboard pattern with a seeded random shift. The four masks of each scale are each applied once, so every token is masked exactly once per scale across $|\mathcal{S}| \times 4 = 8$ masked forwards, and each position is predicted twice, at two granularities. The checkerboard keeps the query length constant across forwards, and the random shift prevents the pattern phase from aligning with object structure while leaving the coverage property unchanged.

**Blind encoding.** For a mask $M$, every masked patch embedding is replaced by the backbone's own learnable mask token before position encoding and before all attention blocks,

$$h \;=\; \mathrm{ViT}_\theta\Big(P(x) \odot (1 - M) \;+\; e_{\mathrm{mask}} \odot M \;+\; p\Big), \tag{4}$$

where $P(x)$ denotes the patch embeddings of the canvas, $e_{\mathrm{mask}}$ is the learnable mask token of the backbone, and $p$ collects the positional embeddings. The context memory $h$ therefore contains no information about the target content.

**Blind query decoder.** A shallow Transformer decoder predicts the masked positions from context alone. The query for position $u$ combines a sinusoidal position encoding, a scale embedding, and the memory token at $u$, so it knows *where* it is without knowing *what* is there,

$$q_u = \mathrm{pe}(u) + e_s + h(u), \qquad H \;=\; \mathrm{TransformerDecoder}\big(\{q_u\},\, h\big). \tag{5}$$

where $\mathrm{pe}(u)$ is the sinusoidal position encoding of $u$, $e_s$ is the embedding of the block scale $s$, and $h(u)$ is the context-memory token at the same position. The decoder produces logits over the category's normal atoms $A_c = \{a_k\}_{k=1}^{K_c}$, prototypes of normal training patches selected by $k$-center covering, through

$$\ell_k(u) \;=\; s_\theta \left\langle \widehat{W_q H(u)},\, \widehat{W_k a_k} \right\rangle, \qquad s_\theta = e^\gamma \text{ (clamped)}, \tag{6}$$

where $W_q$ and $W_k$ are learned projections of the decoder output and of the atom descriptors, the hats denote $L_2$ normalization, and $s_\theta = e^\gamma$ is a single learnable scale. During training the atom library excludes tokens owned by the query image, which extends the blindness discipline to the training objective.

**Conditional free energy.** For the scale-$s$ mask covering position $u$, the clean-pass descriptor $z(u)$ is scored against normal atoms using weights predicted from masked context. Suppressing the scale index on the logits and weights, we write

$$\pi_k(u) = (1 - \varepsilon)\,\mathrm{softmax}_k\big(\ell(u)\big) + \frac{\varepsilon}{K_c}, \qquad E^{(s)}(u) = -\tau \log \sum_{k=1}^{K_c} \pi_k(u) \exp\Big(\frac{\langle z(u),\, a_k\rangle - 1}{\tau}\Big). \tag{7}$$

Here $\tau$ controls the concentration of each similarity kernel, and $\varepsilon / K_c$ gives each atom a minimum mixture weight. For fixed $\tau$ and descriptor dimension, $E^{(s)}(u)$ is a positive affine transformation of the conditional vMF-mixture negative log-likelihood (Appendix B.1). At training time, atoms owned by the query image are excluded, and the weights are renormalized over eligible atoms. At inference, four complementary masks score every position once per scale, and the energies are averaged:

$$E(u) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} E^{(s)}(u). \tag{8}$$

The decoder is trained on normal images by minimizing mean energy over masked targets sampled from this schedule.

**Mode standardization.** Raw energies carry position- and mode-dependent offsets. Each category's support is partitioned into $M$ appearance modes with anchors $\{\rho_j\}$ ($k$-center on descriptors), every

position is assigned to a mode $m(u) = \arg\max_j \langle z(u),\, \rho_j \rangle$, and the energy is standardized with shrinkage moments estimated on a normal validation split,

$$z_{\mathrm{V}}(u) = \frac{E(u) - \mu_{c,m(u)}}{\sigma_{c,m(u)}}, \quad \mu_{c,m} = \frac{N_{c,m}\hat{\mu}_{c,m} + \kappa\,\hat{\mu}_c}{N_{c,m} + \kappa}, \quad \sigma_{c,m} = \max\big(\hat{\sigma}_{c,m},\, \varrho\,\hat{\sigma}_c\big), \tag{9}$$

where the hats denote empirical moments estimated on the validation split, $N_{c,m}$ counts the validation positions assigned to mode $m$ of category $c$, $\hat{\mu}_c$ and $\hat{\sigma}_c$ are the corresponding moments pooled over all modes of the category, $\kappa$ is a shrinkage prior count, and $\varrho$ is a floor fraction for the standard deviation. The result $z_{\mathrm{V}} \in \mathbb{R}^{64\times 64}$ is the first evidence field. At test time all statistics are frozen.

### 3.3 Witness evidence from a coverage-maximizing normal library

The contextual energy answers a local question: is the appearance at this position plausible given its surroundings? It cannot answer a global question: does this appearance occur anywhere in the normal world? A scratch on a screw is locally self-consistent, yet absent from every normal screw. The witness evidence is designed for this global question, and it is deliberately training-free: all of its knowledge comes from an explicit library of normal images.

**Coverage-maximizing reference library.** For each category $c$ we select $R$ normal training images by deterministic farthest-first ($k$-center) selection (Gonzalez, 1985) on the semantic descriptors $s(x)$ of Eq. 2, maximizing the covering radius of the normal appearance space, with the budget capped at the eligible training images. The library is frozen and stores every patch descriptor of each selected image.

**Witness distance field.** For each of the $P$ whole-view patches we compute the distance to the closest normal patch in the entire library,

$$d_x(u) \;=\; \max\Big(0,\; 1 - \max_{r\in R_c} \langle z_x(u),\, r\rangle \Big), \tag{10}$$

where $z_x(u)$ is the descriptor of the whole-view patch at position $u$. A normal patch finds a nearly identical reference, so $d \approx 0$. An unsupported pattern stays far from every reference.

**CVaR tail readout.** An image score must be sensitive to small defects yet robust to isolated noisy positions. We use the tail conditional expectation of the distance field,

$$s(x) \;=\; \mathrm{CVaR}_\alpha\big[u \mapsto d_x(u)\big] \;=\; \frac{1}{m}\sum_{i=1}^{m} d_{(i)}(x), \tag{11}$$

where $d_{(1)} \geq d_{(2)} \geq \cdots$ are the sorted distances, $\alpha$ is the tail fraction, and $m = \lceil \alpha P \rceil$ (Rockafellar & Uryasev, 2000). The ablations show that both limits, the maximum and the global mean, are consistently inferior.

**Route-conditional conformal calibration.** Scores are converted into $p$-values against the independent normal calibration split of the image's route,

$$\hat{p}(x) \;=\; \frac{1 + \#\big\{i :\; s(x_i) \geq s(x)\big\}}{N_{c(x)} + 1}, \tag{12}$$

the add-one rule (Vovk et al., 2005), where $N_{c(x)}$ is the number of calibration images of the route and $\#$ counts the calibration scores that exceed $s(x)$. The rule is finite-sample valid under exchangeability, in the sense that $\mathbb{P}\{\hat{p} \leq \alpha'\} \leq \alpha'$ for any level $\alpha'$. The image decision is $\hat{p}(x) \leq \alpha'$. Because $c(x)$ is constant within a category in practice, per-route calibration equals per-class calibration with misrouting robustness.

### 3.4 Calibrated joint evidence

The witness branch computes a full spatial field $d(u)$ but consumes it only through the scalar readout of Eq. 11. TRIAD recovers the field as a second localization evidence, calibrated so that it is commensurable with $z_{\mathrm{V}}$.

**Leakage-safe calibration.** Two steps use only normal splits. First, route-scalar null moments. For each route $r$, with $r(x) = c(x)$, we pool the distances of all calibration images over all $P$ positions,

$$\hat{\mu}_r = \frac{1}{N_r P} \sum_i \sum_u d_{x_i}(u), \qquad \hat{\sigma}_r = \sqrt{\mathrm{Var}\left(\{d_{x_i}(u)\}\right)}, \qquad \tilde{z_{\mathrm{G}}}(u) = \frac{d_x(u) - \hat{\mu}_{r(x)}}{\hat{\sigma}_{r(x)}}. \tag{13}$$

Pooling thousands of positions per route, with $N_r$ the number of calibration images of route $r$, makes the null robust where per-position moments, with only a few dozen samples each, overfit and inflate. Second, validation self-standardization. The scale of $\tilde{z_{\mathrm{G}}}$ on an independent normal validation split corrects the mild inflation of estimating $\hat{\sigma}_r$ from finite data,

$$s_{\mathrm{val}} = \mathrm{Std}\left(\{\tilde{z_{\mathrm{G}}}(u) : x \in \mathcal{D}_{\mathrm{val}},\, u\}\right), \qquad z_{\mathrm{G}}(u) = \tilde{z_{\mathrm{G}}}(u) \,/\, s_{\mathrm{val}}, \tag{14}$$

and the result is bilinearly upsampled to align with $z_{\mathrm{V}}$.

**Parameter-free fusion.** Both calibrated fields have, by construction, approximately unit variance under the null. For a bivariate Gaussian working model with correlation $\rho$, the likelihood-ratio-optimal *linear* statistic for a dense shift ($\mathbb{E}[z_{\mathrm{V}}] = \mathbb{E}[z_{\mathrm{G}}] = \delta > 0$) is the precision-weighted sum; equal calibration variances make equal weights the exact optimum,

$$J(u) \;=\; \tfrac{1}{2}\Big(z_{\mathrm{V}}(u) + z_{\mathrm{G}}(u)\Big), \tag{15}$$

$$t^*(u) \;\propto\; \mathbf{1}^\top \Sigma^{-1} \begin{pmatrix} z_{\mathrm{V}}(u) \\ z_{\mathrm{G}}(u) \end{pmatrix} \;=\; \frac{z_{\mathrm{V}}(u) + z_{\mathrm{G}}(u)}{1+\rho}, \tag{16}$$

and $1/(1{+}\rho)$ is a per-route constant that does not affect within-route ranking (Proposition 1, Appendix B.2). Max-combination would suit a *sparse* alternative that shifts one channel only; pixel-level defects in our data raise both channels together, and the empirical comparison confirms the sum. The pixel anomaly map applies the frozen path, bilinear upsampling followed by Gaussian smoothing, to $J$. The image decision remains identical to the witness detector of Eqs. 11 and 12, so fusion can only improve localization and never perturbs detection.

# 4 Experiments

## 4.1 Experimental setup

**Datasets and metrics.** We evaluate on MVTec-AD (Bergmann et al., 2019) with 15 categories and VisA (Zou et al., 2022) with 12 categories under the unified multi-class protocol. We report seven standard metrics: image-level AUROC, AP, and $F_1$-max; pixel-level AUROC, AP, $F_1$-max; and AUPRO. For our method and its ablations, all pixel-level metrics are computed on the $448 \times 448$ anomaly maps produced by the frozen pixel path, so every component of our method is evaluated at one shared resolution.

**Protocol and baselines.** Baseline numbers are quoted as reported in the MaskAD paper (Lu et al., 2026), which evaluates the same unified protocol and represents the current state of the art. Our results are means and standard deviations over three source seeds, and every per-seed run re-fits its own decoder, atoms, references, modes, and calibration statistics end-to-end. Baselines train their backbones; TRIAD uses a frozen DINOv2 ViT-B/14, which reduces training to a shallow decoder (Appendix C).

**Statistical discipline.** All design choices were frozen before any test result was read, and unlabelled evidence was dumped to disk before labels were opened. All ablations are preregistered single-factor experiments that differ from the deployed configuration in exactly one factor. Each suite contains a bit-exact reproduction anchor of the deployed pipeline, which must reproduce the frozen metrics with a drift below $0.10$ pp, where pp denotes percentage points on the reported metric, before any rival row is interpreted.

## 4.2 Main results

Table 1 shows that TRIAD improves the unified state of the art on every metric on both datasets. The three evidences form a strict ladder on the pixel side. On MVTec-AD, the contextual energy

Table 1: Unified multi-class anomaly detection on MVTec-AD and VisA. Baseline numbers are as reported in Lu et al. (2026) under the same protocol; "–" not reported. TRIAD (ours): mean±std over three seeds. Best in **bold**.

| | | Image-level | | | Pixel-level | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Method | AUROC | AP | $F_1$-max | AUROC | AP | $F_1$-max | AUPRO |
| MVTec-AD | RD4AD (Deng & Li, 2022) | 94.6 | 96.5 | 95.2 | 96.1 | 48.6 | 53.8 | 91.1 |
| | DeSTSeg (Zhang et al., 2023c) | 89.2 | 95.5 | 91.6 | 93.1 | 54.3 | 50.9 | 64.8 |
| | UniAD (You et al., 2022) | 96.5 | 98.8 | 96.2 | 96.8 | 43.4 | 49.5 | 90.7 |
| | DiAD (He et al., 2024b) | 97.2 | 99.0 | 96.5 | 96.8 | 52.6 | 55.5 | 90.7 |
| | ViTAD (Zhang et al., 2023b) | 98.3 | 99.4 | 97.3 | 97.7 | 55.3 | 58.7 | 91.4 |
| | CNC (Wang et al., 2025) | 98.6 | 99.3 | – | 98.0 | 56.4 | – | 93.0 |
| | MambaAD (He et al., 2024a) | 98.6 | 99.6 | 97.8 | 97.7 | 56.3 | 59.2 | 93.1 |
| | MaskAD (Lu et al., 2026) | 98.6 | 99.3 | 97.8 | 97.7 | 59.5 | 58.9 | 94.3 |
| | **TRIAD (ours)** | **99.29±0.05** | **99.74±0.02** | **98.75±0.19** | **98.49±0.03** | **68.25±0.20** | **67.70±0.15** | **96.61±0.06** |
| VisA | RD4AD (Deng & Li, 2022) | 92.4 | 92.4 | 89.6 | 98.1 | 38.0 | 42.6 | 91.8 |
| | DeSTSeg (Zhang et al., 2023c) | 88.9 | 89.0 | 85.2 | 96.1 | 39.6 | 43.0 | 67.4 |
| | UniAD (You et al., 2022) | 88.8 | 90.8 | 85.8 | 98.3 | 33.7 | 39.0 | 85.5 |
| | DiAD (He et al., 2024b) | 86.8 | 88.3 | 85.1 | 96.0 | 26.1 | 33.0 | 75.2 |
| | ViTAD (Zhang et al., 2023b) | 90.5 | 91.7 | 86.3 | 98.2 | 36.6 | 41.1 | 85.1 |
| | CNC (Wang et al., 2025) | 93.2 | 92.6 | – | 98.5 | 37.8 | – | 91.4 |
| | MambaAD (He et al., 2024a) | 94.3 | 94.5 | 89.4 | 98.5 | 39.4 | 44.0 | 91.0 |
| | MaskAD (Lu et al., 2026) | 94.5 | 94.8 | 90.5 | 98.8 | 41.9 | 44.4 | 92.7 |
| | **TRIAD (ours)** | **95.06±0.05** | **95.74±0.10** | **91.22±0.26** | **98.97±0.02** | **46.61±0.18** | **50.48±0.23** | **94.57±0.02** |

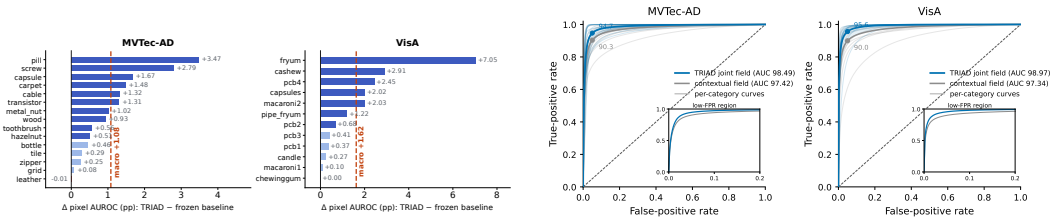


Figure 2: Left: per-category pixel-AUROC gains of the calibrated fusion (Eq. 15) over the contextual field alone, mean over three seeds; 25 of the 27 categories improve and the two remaining cases are exact ties. Right: pixel-level macro ROC curves of the contextual field alone and the full TRIAD model, averaged over the categories of each dataset; thin lines show the per-category curves, the shaded area marks the gain of the calibrated fusion, and the inset magnifies the low-FPR region, with round markers at the deployed false-alarm level $\alpha'=0.05$. Macro AUROC values are reported in Table 1.

alone reaches 63.36 pixel AP, adding the witness field raises this to 64.90, and the calibrated fusion of the two reaches 68.25, which is 8.8 points above the strongest published baseline; the same ladder holds on VisA, with 40.86, 42.99, and 46.61. The image level was already strong and stays exact: the witness detector reaches 99.29/95.06 image AUROC, and fusion leaves these numbers bit-identical by design, so the localization gain carries no detection regression risk. The gains are also concentrated where it matters: AUPRO reaches 96.61/94.57, and the per-category breakdown in Figure 2 (left) shows that 25 of the 27 categories improve while the remaining two are exact ties. The largest gains occur in small-defect categories such as fryum (+7.05), pill (+3.47), cashew (+2.91), and screw (+2.79). The pixel-level ROC curves in Figure 2 (right), averaged over the categories of each dataset, show that the calibrated fusion shifts the curve toward the upper-left corner across the entire operating range, exactly the behavior expected from a second independent evidence channel. Per-category tables for all metrics are given in Appendix F.

### 4.3 Ablation Studies

We run three preregistered single-factor suites. Every variant differs from the deployed configuration in exactly one factor, all variants share the same frozen features, and each suite contains a bit-exact reproduction anchor of the deployed pipeline. Figure 10 in Appendix H summarizes the three suites.

**Target blindness and context correspondence.** Table 2 replaces, one factor at a time, the blind encoding of the contextual energy branch. A reconstruction-style leaky context, with clean memory,

Table 2: Ablation of target blindness and context correspondence in the contextual energy branch (pixel level, mean±std over three seeds). The deployed design is the blind encoding. Leaky removes target blindness at inference, and shuffled replaces the context by a mismatched one from the same category.

| | AUROC | AP | $F_1$-max | AUPRO |
|---|---|---|---|---|
| *MVTec-AD* | | | | |
| blind (deployed) | **97.42±0.09** | **63.36±0.24** | **64.48±0.09** | **94.51±0.21** |
| leaky context | 96.68±0.10 | 59.22±0.03 | 60.56±0.09 | 92.43±0.35 |
| shuffled context | 96.70±0.07 | 59.89±0.32 | 60.93±0.22 | 92.28±0.19 |
| *VisA* | | | | |
| blind (deployed) | **97.34±0.09** | **40.86±0.61** | **44.77±0.44** | **90.39±0.15** |
| leaky context | 96.29±0.17 | 36.75±0.76 | 42.00±0.49 | 83.58±0.82 |
| shuffled context | 95.81±0.04 | 31.84±0.74 | 37.14±0.66 | 80.68±0.70 |

Table 3: Ablation of the witness evidence branch (image level, mean±std over three seeds). Each row changes exactly one factor relative to the deployed configuration (top).

| | MVTec-AD | | | VisA | | |
|---|---|---|---|---|---|---|
| Variant | AUROC | AP | $F_1$-max | AUROC | AP | $F_1$-max |
| deployed ($\text{CVaR}_{1\%}$, NN, $k$-center $R$=64) | **99.29±0.05** | **99.74±0.02** | **98.75±0.19** | **95.06±0.05** | **95.74±0.10** | **91.22±0.26** |
| readout: max ($\alpha \to 0$) | 98.13±0.05 | 99.16±0.05 | 97.64±0.11 | 94.53±0.19 | 95.20±0.21 | 90.60±0.05 |
| readout: mean ($\alpha = 1$) | 93.43±0.11 | 97.02±0.07 | 94.35±0.25 | 87.73±0.54 | 89.21±0.33 | 85.02±0.44 |
| readout: $\text{CVaR}_{10\%}$ | 98.59±0.12 | 99.52±0.04 | 97.80±0.06 | 91.28±0.20 | 92.10±0.20 | 87.92±0.18 |
| distance: mean similarity | 84.79±0.04 | 92.68±0.01 | 89.22±0.06 | 79.10±0.04 | 82.74±0.04 | 78.87±0.11 |
| library: $k$-center $R$=16 | 98.34±0.09 | 99.20±0.12 | 97.65±0.02 | 92.62±0.41 | 93.25±0.37 | 88.70±0.46 |
| library: random $R$=64 | 99.13±0.19 | 99.57±0.24 | 98.73±0.08 | 94.01±0.11 | 94.16±0.08 | 90.43±0.23 |
| library: prototype $R$=64 | 97.90±0.08 | 98.81±0.16 | 98.08±0.03 | 92.41±0.28 | 92.38±0.23 | 89.29±0.30 |

an unmasked target, and a query containing the target projection, degrades pixel AUROC by $0.74$ (MVTec-AD) and $1.05$ (VisA) and pixel AP by $4.14$/$4.11$. Replacing the corresponding context by a mismatched context from another image of the same category degrades AUROC by $0.72$/$1.53$ and AP by $3.47$/$9.02$. Paired Wilcoxon tests over category and seed pairs give $p \leq 3.1 \times 10^{-11}$ for AUROC and $p \leq 5.7 \times 10^{-14}$ for AP under leakage, and $p \leq 4.1 \times 10^{-4}$ / $1.9 \times 10^{-3}$ (MVTec-AD), $p \leq 5.1 \times 10^{-5}$ / $2.9 \times 10^{-11}$ (VisA) for correspondence. The degradation mechanism matches the prediction: leakage lets the query explain the anomaly as normal and collapses the anomaly-to-background contrast. These are inference-side interventions on the frozen decoder; a leaky-trained counterpart would require retraining and is left to future work.

**Score construction in the witness branch.** Table 3 varies one factor at a time around the deployed witness configuration. Replacing nearest-neighbour matching by mean similarity to the library collapses image AUROC by $14.5$ (MVTec-AD) and $16.0$ (VisA), because averaging over the whole library washes out exactly the contrast that the witness is meant to detect. Widening the tail from $1\%$ to $10\%$ costs $0.7$/$3.8$ points, and the global mean costs $5.9$/$7.3$, so the $\text{CVaR}_{1\%}$ readout is the sweet spot between dilution of small defects and sensitivity to isolated noise. A random library is statistically indistinguishable on the near-saturated MVTec-AD image level but significantly worse on VisA ($-1.06$, $p$=$6.5 \times 10^{-5}$), which vindicates coverage-based selection on the more heterogeneous benchmark. Route-conditional calibration is evaluated at the decision level: pooling the null across routes increases the mean per-category deviation of the empirical false-positive rate from the nominal level, from $0.070$ to $0.089$ on MVTec-AD and from $0.028$ to $0.033$ on VisA, while leaving ranking metrics unchanged, exactly as theory predicts for a monotone within-route transform.

**Calibrated fusion.** Table 4 isolates the third contribution on the pixel side. The witness field alone already localizes well, at $98.25$ AUROC on MVTec-AD, and adding it to the contextual field yields the deployed $98.49$/$68.25$; the gain over witness-only is $+0.24$/$+0.16$ AUROC and $+3.35$/$+3.62$ AP. The two evidences are complementary rather than redundant. A natural alternative calibration, with shrunk moments per route and per position, drops AP by $4.23$/$3.03$ and AUROC by $0.42$/$0.18$, reproducing the small-sample overdispersion pathology that the route-scalar null with $s_{\text{val}}$ is designed to avoid.

Table 4: Ablation of the calibrated joint evidence (pixel level, mean±std over three seeds). Image-level metrics are frozen in all rows.

| | MVTec-AD | | | | VisA | | | |
|---|---|---|---|---|---|---|---|---|
| Variant | AUROC | AP | $F_1$-max | AUPRO | AUROC | AP | $F_1$-max | AUPRO |
| contextual field only ($z_V$) | 97.42±0.09 | 63.36±0.24 | 64.48±0.09 | 94.51±0.21 | 97.34±0.09 | 40.86±0.61 | 44.77±0.44 | 90.39±0.15 |
| witness field only ($z_G$) | 98.25±0.00 | 64.90±0.05 | 65.01±0.06 | 95.91±0.02 | 98.81±0.01 | 42.99±0.09 | 47.85±0.19 | 92.98±0.10 |
| **TRIAD joint (deployed)** | **98.49±0.03** | **68.25±0.20** | **67.70±0.15** | **96.61±0.06** | **98.97±0.02** | **46.61±0.18** | **50.48±0.23** | **94.57±0.02** |
| joint, per-position nulls | 98.07±0.03 | 64.02±0.10 | 66.08±0.11 | 95.79±0.07 | 98.79±0.02 | 43.58±0.11 | 48.18±0.06 | 94.43±0.06 |

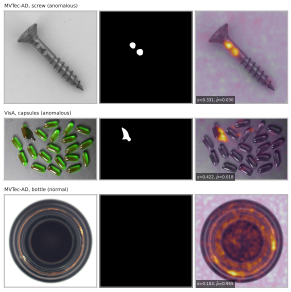

Figure 3: Qualitative results of the deployed TRIAD. Top and middle: anomalous test images of screw and capsules with the highest witness scores, their ground-truth masks, and the overlaid pixel anomaly maps. Bottom: the normal bottle with the lowest witness score. Overlays are min–max normalized for display.

**Masking schedule.** Single-scale variants are consistently but mildly worse than the multi-scale schedule (pixel AUROC 97.41/97.39 vs 97.42 on MVTec-AD), as the two granularities trade small advantages on different categories (Appendix H).

### 4.4 Qualitative results

Figure 3 illustrates the deployed model: on the highest-scoring anomalous images of screw and capsules the overlays concentrate precisely on the defects, while on the lowest-scoring normal bottle no salient activation appears, and the $p$-values confirm both decisions. Appendix G provides the complete anomaly maps.

### 4.5 Model complexity and runtime

TRIAD trains only a shallow decoder on a frozen backbone. Across the unified baselines, TRIAD has the smallest trainable budget: the 10.11M-parameter blind decoder, about 10.5% of all model parameters, while attaining the highest mean metric score ($m$AD 89.8, against 86.6 for MaskAD and 81.2–86.0 for the others). At inference the witness branch adds only 7.9 ms per test image, under 4% of the total, so the second evidence channel is nearly free; Table 7 in Appendix C.5 gives the full accounting with hardware caveats.

## 5 Conclusion and limitations

**Limitations.** The optimality of Eq. 15 assumes a Gaussian working model rather than a guarantee, and token-level evidence limits localization to the patch resolution. The blindness ablation is inference-side, and a fully leaky-trained counterpart remains to be measured.

**Conclusion.** We presented TRIAD, a tri-evidence anomaly detector on a fully frozen DINOv2. The target-blind contextual energy never lets the scorer see the target. The witness branch grounds scores in the normal world with distribution-free image-level guarantees. The calibrated fusion turns the previously discarded witness field into a second localization evidence. The framework advances the state of the art on all seven metrics of both benchmarks, with every design decision validated by preregistered single-factor ablations.

### AI USE STATEMENT

In this work, we used generative AI tools for refactoring and debugging parts of the experiment and analysis code, for language polishing and formatting of the manuscript, and for typesetting figures and tables. We have not used generative AI tools for generating research ideas, hypotheses, experimental designs, or numerical results, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for language polishing of the manuscript. We have reviewed all AI-assisted work. Every quantitative claim was checked manually against run artifacts produced by the released pipeline under fixed seeds and integrity audits. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses only public industrial inspection benchmarks, MVTec-AD and VisA, which contain no personal data, and all models train exclusively on normal images. The intended application is industrial quality inspection. As with anomaly detection methods in general, the framework could in principle be repurposed for other monitoring scenarios. The released artifacts operate on product imagery, provide no identity information, and introduce no privacy, security, or legal concerns beyond those of the underlying public datasets.

### REPRODUCIBILITY STATEMENT

All experiments use public datasets and a public frozen backbone. The paper is accompanied by a full implementation with fixed seeds, including core modules, run scripts, unit tests, and per-run integrity audits. Hyperparameters and all constants are listed in Appendix C, the complete training and inference procedures in Appendix D.

## REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592. IEEE, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

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

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

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024a.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhisheng Chen, Yabiao Wang, Chengjie Wang, Yong Liu, and Lei Xie. DiAD: A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12503–12511, 2024b.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8472–8480, 2024c.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616. IEEE, 2023.

Ruiying Lu, Gang Liu, Kang Li, Long Tian, and Junwei Zhang. Maskad: Parallel masked autoencoder for multi-class unsupervised anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15457–15465, 2026.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 2000.

Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 73–108. Springer, 1992.

Marc Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

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

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356, 2002.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14298–14308. IEEE, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Xiaolei Wang, Xiaoyang Wang, Huihui Bai, Eng Gee Lim, and Jimin Xiao. Cnc: Cross-modal normality constraint for unsupervised multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 7943–7951, 2025.

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

Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2312.07495*, 2023a.

Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain ViT reconstruction for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2312.07495*, 2023b.

Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3914–3923. IEEE, 2023c.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408, 2022.

## CONTENTS OF THE APPENDIX

A Overview, datasets, and baselines . . . 13
A.1 Dataset statistics . . . 13
A.2 Baseline details . . . 14
B Analytical derivations . . . 14
B.1 Free energy as a von Mises–Fisher mixture likelihood . . . 14
B.2 Optimality of the equal-weight fusion . . . 15
C Implementation details and hyperparameters . . . 15
C.1 Backbone and features . . . 15
C.2 Blind decoder training . . . 15
C.3 Constants . . . 15
C.4 Statistics . . . 16
C.5 Complexity accounting . . . 16
D Complete algorithm pseudocode . . . 16
E Score separation and calibration validity . . . 16
F Per-category results . . . 17
G Qualitative anomaly maps . . . 17
H Additional ablations and falsified directions . . . 19
H.1 Ablation summary figure . . . 19
H.2 Entropy conditioning (falsified) . . . 19
H.3 Image-level fusion (falsified) . . . 19
H.4 Single-scale masking schedules . . . 24

## A OVERVIEW, DATASETS, AND BASELINES

This appendix is organized as follows. Section B gives the analytical derivations of the free energy and of the fusion optimality. Section C lists implementation details and hyperparameters. Section D provides complete pseudocode of the offline and inference procedures. Section F reports per-category results. Section G shows qualitative anomaly maps on every category. Section H collects additional ablations.

### A.1 DATASET STATISTICS

Table 5 summarizes the two benchmarks under our normal-only split protocol. The official normal training images of each dataset are split into training, validation, and calibration subsets with fractions $0.8$, $0.1$, and $0.1$. The validation and calibration subsets are sealed from all library fitting and from decoder training. The test sets are the official test splits, which contain both normal and anomalous images, and the anomalous images carry pixel-level ground-truth masks.

Table 5: Dataset statistics under the unified multi-class protocol. The normal-split columns follow the $0.8/0.1/0.1$ protocol of Section 4.5.

| | | Normal-split protocol | | | Test images | |
|---|---|---|---|---|---|---|
| Dataset | Categories | Train | Val | Cal | Normal | Anomalous |
| MVTec-AD (Bergmann et al., 2019) | 15 | 2905 | 362 | 362 | 467 | 1258 |
| VisA (Zou et al., 2022) | 12 | 6931 | 864 | 864 | 962 | 1200 |

**Categories and characteristics.** MVTec-AD contains fifteen categories in two groups. Ten object categories (bottle, cable, capsule, hazelnut, metal nut, pill, screw, toothbrush, transistor, and zipper) show defects such as contamination, structural anomalies, and missing components. Five texture categories (carpet, grid, leather, tile, and wood) show defects such as color stains, cuts, cracks, and holes. Images are up to $1024^2$ pixels with 73 distinct defect types in total, and each category contains a wide range of defect sizes, which makes pixel AP particularly demanding. VisA contains twelve categories (candle, capsules, cashew, chewing gum, fryum, macaroni1, macaroni2, pcb1,

pcb2, pcb3, pcb4, and pipe fryum) with 10,821 images in total. Compared with MVTec-AD, VisA scenes are more complex: multiple instances appear in one image, viewing angles vary, and many defects are small, low-contrast, or structural (for example, broken pieces and missing holes), which explains the lower absolute pixel-AP values and the larger gains of the witness evidence on this benchmark.

### A.2 Baseline details

Table 6 lists the compared methods with their venues and core mechanisms. All baselines train their backbones, whereas TRIAD keeps the backbone frozen and trains only the blind decoder. Baseline numbers in the main text are quoted from the MaskAD evaluation under the same unified protocol.

Table 6: Compared methods with venues and core mechanisms.

| Method | Venue | Core mechanism |
|---|---|---|
| RD4AD (Deng & Li, 2022) | CVPR 2022 | reconstruction with feature and gradient residuals |
| DeSTSeg (Zhang et al., 2023c) | CVPR 2023 | denoising student–teacher segmentation |
| UniAD (You et al., 2022) | NeurIPS 2022 | reconstruction transformer with neighbor-masked attention |
| DiAD (He et al., 2024c) | AAAI 2024 | latent diffusion reconstruction |
| ViTAD (Zhang et al., 2023a) | 2023 | plain ViT masked reconstruction |
| MambaAD (He et al., 2024a) | NeurIPS 2024 | state-space multi-scale reconstruction |
| CNC (Wang et al., 2025) | AAAI 2025 | cross-modal normality constraint |
| MaskAD (Lu et al., 2026) | AAAI 2026 | masked-image-modeling reconstruction |

## B Analytical derivations

### B.1 Free energy as a von Mises–Fisher mixture likelihood

Let $\mathcal{C}_u$ denote the information supplied to the decoder when scoring position $u$: the masked-view memory, the masked-position queries (including $q_u$), and the mask scale. Fix the inferred category $c$ and its eligible unit-normalized atom bank $A_c = \{a_k\}_{k=1}^{K_c}$. The decoder produces normalized, context-dependent weights $\pi_k(u)$ as in Eq. 7. These weights define a mixture density for a unit-normalized descriptor $z \in \mathbb{S}^{d-1}$:

$$p(z \mid \mathcal{C}_u, A_c) = \sum_{k=1}^{K_c} \pi_k(u)\, \mathrm{vMF}(z; a_k, 1/\tau), \qquad \mathrm{vMF}(z; a, \kappa) = C_d(\kappa) \exp\big(\kappa \langle z, a \rangle\big), \tag{17}$$

where $C_d(\kappa)$ is the normalizing constant and all components share concentration $1/\tau$ (Mardia & Jupp, 2000). For the observed descriptor $z(u)$, Eq. 7 gives

$$\begin{aligned} p\big(z(u) \mid \mathcal{C}_u, A_c\big) &= C_d(1/\tau) e^{1/\tau} \sum_{k=1}^{K_c} \pi_k(u) \exp\left(\frac{\langle z(u),\, a_k \rangle - 1}{\tau}\right) \\ &= C_d(1/\tau) e^{1/\tau} \exp\left(-\frac{E(u)}{\tau}\right). \end{aligned} \tag{18}$$

Consequently,

$$-\log p\big(z(u) \mid \mathcal{C}_u, A_c\big) = \frac{E(u)}{\tau} - \log C_d(1/\tau) - \frac{1}{\tau}. \tag{19}$$

Thus, for fixed $\tau$ and descriptor dimension $d$, the implemented energy is a positive affine transformation of the conditional negative log-likelihood, rather than the negative log-likelihood itself. Minimizing the mean energy over normal training targets and masking contexts is therefore equivalent to maximizing their mean conditional log-likelihood. This density concerns unit-normalized feature descriptors, not raw image pixels. The uniform component in Eq. 7 gives each eligible atom a weight of at least $\varepsilon/K_c$ at inference. During training, atoms owned by the query image are excluded; the mixture and its uniform component are renormalized over the remaining eligible atoms, and the same derivation applies.

### B.2 Optimality of the equal-weight fusion

**Proposition 1** (Equal-weight test under a common Gaussian mean shift)**.** *At a fixed route and position, let $X = (z_{\mathrm{V}}, z_{\mathrm{G}})^{\top}$. Suppose that normal and anomalous scores follow, respectively,*

$$H_0 : X \sim \mathcal{N}(0, \Sigma), \qquad H_1 : X \sim \mathcal{N}(\delta \mathbf{1}, \Sigma), \qquad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \delta > 0, \quad |\rho| < 1. \tag{20}$$

*The likelihood-ratio test for these two distributions is monotone in*

$$t^* = \mathbf{1}^{\top} \Sigma^{-1} X = \frac{z_{\mathrm{V}} + z_{\mathrm{G}}}{1 + \rho}. \tag{21}$$

*Hence the equal-weight score $J = (z_{\mathrm{V}} + z_{\mathrm{G}})/2$ of Eq. 15 gives the same within-route ranking and, with an appropriate threshold, the same most-powerful test at a fixed false-positive rate.*

*Proof.* The distributions under $H_0$ and $H_1$ have the same positive-definite covariance. Their log-likelihood ratio is

$$\begin{aligned} \Lambda(X) &= \delta \, \mathbf{1}^{\top} \Sigma^{-1} X - \frac{\delta^2}{2} \, \mathbf{1}^{\top} \Sigma^{-1} \mathbf{1} \\ &= \frac{\delta}{1 + \rho} (z_{\mathrm{V}} + z_{\mathrm{G}}) - \frac{\delta^2}{1 + \rho} = \frac{2\delta}{1 + \rho} J - \frac{\delta^2}{1 + \rho}. \end{aligned} \tag{22}$$

Because $\delta > 0$ and $1 + \rho > 0$, $\Lambda(X)$ is strictly increasing in $J$. The Neyman–Pearson lemma therefore establishes the claim for the stated Gaussian hypotheses (Neyman & Pearson, 1992). □

The proposition describes a working model, not a distribution-free guarantee for pixel localization. Normal-only standardization aims to make the two evidence scales comparable, but does not establish joint Gaussianity, equal anomalous mean shifts, or unchanged covariance. For a common-covariance Gaussian alternative with a general mean shift $\Delta$, the likelihood-ratio-optimal linear direction is $\Sigma^{-1}\Delta$; inverse-variance weights follow only in the special case of a diagonal covariance and equal coordinate shifts. Likewise, taking the maximum of the two scores can be a useful heuristic for a one-channel anomaly, but is not generally its exact likelihood-ratio statistic. The ablations in Section 4.3 provide empirical evidence for the chosen equal-weight fusion on the evaluated benchmarks; they do not prove that every pixel or category benefits from fusion.

## C Implementation details and hyperparameters

### C.1 Backbone and features

DINOv2 ViT-B/14 (`vit_base_patch14_dinov2.lvd142m`), frozen throughout. Features from layers $\{6, 9, 12\}$. Unit descriptors of dimension 2304 (three $L_2$-normalized layers, concatenated). Canvas $896^2$ composed of four $448^2$ views. Token canvas $64 \times 64$ and whole-view grid $32 \times 32$ with $P{=}1024$ patches. Routing and reference selection use the $L_2$-normalized mean of the last-layer tokens, a 768-dimensional semantic descriptor; the atom bank, the witness matching, and the mode anchors use the 2304-dimensional concatenated descriptors.

### C.2 Blind decoder training

Trained on normal training images only, minimizing the multi-scale energy of Eq. 7. AdamW with learning rate $2 \times 10^{-4}$ decaying to $10^{-5}$, weight decay 0.05, batch size 4 with gradient accumulation 2, gradient-norm clip 1.0, and 80 epochs. Decoder: 4 blocks, 6 heads, width 384, key dimension 128, dropout 0. Masking schedule seed 202610. Atoms exclude the tokens of the query image.

### C.3 Constants

Atom budget $K{=}2048$ per category (capped at the eligible pool). Mode anchors $M{=}64$ per category. Reference budget $R{=}64$ per category (budget-capped). Energy temperature $\tau{=}0.02$. Smoothing mass $\varepsilon{=}0.05$. Shrinkage prior count $\kappa{=}256$ with standard deviation floor $0.1\,\hat{\sigma}_c$. CVaR tail

fraction $\alpha$=0.01, so $m$=11 of 1024 positions. Conformal decision level $\alpha'$=0.05. Calibration fraction 0.1 and validation fraction 0.1 of the normal split. Pixel map: bilinear upsampling to 448, then Gaussian smoothing with $\sigma$=4. Routing centroids from the normal training split. All pixel-level metrics are computed on these $448 \times 448$ anomaly maps, so every component of TRIAD is evaluated at one shared resolution.

### C.4 Statistics

Route-scalar nulls pool $N_r \times 1024$ calibration positions per route. $s_{\text{val}}$ comes from the held-out validation split and equals 1.01 to 1.02 in practice. Three source seeds $(0, 1, 2)$; each seed re-fits decoder, atoms, references, modes, nulls, and $s_{\text{val}}$ end-to-end.

### C.5 Complexity accounting

Table 7 extends the total parameter count, the FLOPs per test image, the training time, and the frame rate. Baseline entries are the values reported in Lu et al. (2026), measured on the hardware of the respective authors. Our entries were measured on one NVIDIA RTX PRO 6000 GPU at the native $896 \times 896$ canvas, after warm-up, over 307 official MVTec-AD test images; the training time covers library fitting plus 80 epochs of decoder training. Because the hardware differs, the training time and the frame rate are not directly comparable across rows. The FLOPs of TRIAD are dominated by the nine frozen-encoder forwards of the multi-scale blind probing over the full token canvas.

Table 7: Full complexity comparison on MVTec-AD. Baseline rows are as reported in Lu et al. (2026). Entries marked with a dagger were measured on our hardware and are not directly comparable across GPUs.

| Method | Params (M) | FLOPs (G) | Train time (h) | FPS | $m$AD |
|---|---|---|---|---|---|
| RD4AD (Deng & Li, 2022) | 80.6 | 28.4 | 4.1 | **90.6** | 82.3 |
| UniAD (You et al., 2022) | **24.5** | **3.6** | 13.4 | 56.8 | 81.7 |
| DeSTSeg (Zhang et al., 2023c) | 35.2 | 122.7 | 2.5 | 74.7 | 81.2 |
| DiAD (He et al., 2024b) | 1331.3 | 451.5 | – | – | 84.0 |
| MambaAD (He et al., 2024a) | 25.7 | 8.3 | 5.7 | 20.9 | 86.0 |
| MaskAD (Lu et al., 2026) | 28.5 | 38.5 | **2.2** | 41.5 | 86.6 |
| TRIAD (ours) | 96.4 (10.1 trainable) | 8585.8$^\dagger$ | $\approx$3.0$^\dagger$ | 4.5$^\dagger$ | **89.8** |

## D Complete algorithm pseudocode

Algorithms 1 and 2 specify TRIAD end-to-end. The offline stage has three strictly ordered phases. Phase A fits the per-category libraries of both evidence branches on the train split only, with no gradients. Phase B is the only gradient stage, the blind decoder training, with owner exclusion as training-time blinding and validation-based checkpoint selection. Phase C is leakage-safe calibration, in which the validation and calibration splits, sealed from every library and from training, are consumed once to produce the frozen statistics. The inference stage is a single frozen sweep: routing first, then the three evidence computations, with the image decision kept bit-identical to the witness branch. Equation references point to the main text.

## E Score separation and calibration validity

Figure 4 shows the distribution of the witness image score $s(x)$ of Eq. 11 for normal and anomalous test images on both datasets. The two distributions are well separated, which underlies the image-level results in Table 1. Figure 5 evaluates the conformal calibration of Eq. 12 directly. For valid add-one $p$-values, the empirical cumulative distribution of $\hat{p}$ over normal test images should follow the uniform reference, and it does on both datasets. The $p$-values of anomalous images concentrate near zero, which is the detection power that complements this validity.

**Algorithm 1** TRIAD offline stage (normal images only): library fitting (Phase A), blind decoder training (Phase B), and leakage-safe calibration (Phase C)

**Require:** normal set $\mathcal{D}_{\text{norm}}$ with category folders; frozen backbone $\text{ViT}_\theta$; constants $K{=}2048$, $R{=}64$, $M{=}64$, $\tau{=}0.02$, $\varepsilon{=}0.05$, $\kappa{=}256$, scales $\mathcal{S}{=}\{2,4\}$

**Ensure:** frozen libraries $\{(\nu_c,\, A_c,\, \{\rho_j\}_{j=1}^{M},\, R_c)\}_{c=1}^{C}$; trained decoder $\theta_d$; frozen statistics $\{(\mu_{c,m}, \sigma_{c,m})\}$, $\{(\hat{\mu}_r, \hat{\sigma}_r)\}$, $s_{\text{val}}$, per-route null scores $\{s(x_i)\}_{i \le N_r}$

**Phase A: library fitting on the train split only (no gradients)**

$(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{cal}}) \leftarrow$ split of $\mathcal{D}_{\text{norm}}$ with fractions $0.8/0.1/0.1$; seal $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{cal}}$

**for** each category $c = 1, \ldots, C$ **do**

  $\nu_c \leftarrow$ mean of $\{s(x) : x \in \mathcal{D}_{\text{train}}(c)\}$ ▷ Eq. 2: fits the router

  route every train image, $c(x) = \arg\max_k \langle s(x),\, \nu_k \rangle$; eligible set $\mathcal{E}_c = \{x \in \mathcal{D}_{\text{train}}(c) : c(x) = c\}$ ▷ routing-consistency filter

  $A_c \leftarrow \textsc{kCenter}(\{z(t) : x \in \mathcal{E}_c\},\, K)$; record owner images ▷ Eqs. 1–6; budget-capped

  $\{\rho_j\}_{j=1}^{M} \leftarrow \textsc{kCenter}(\{z(t) : x \in \mathcal{E}_c\},\, M)$ ▷ mode anchors for Eq. 9

  $R_c \leftarrow \textsc{kCenter}(\{s(x) : x \in \mathcal{E}_c\},\, R)$; store all $P$ patches per reference ▷ witness library

**end for**

freeze and hash-lock $\{\nu_c, A_c, \{\rho_j\}, R_c\}_{c=1}^{C}$ ▷ both libraries precede training

**Phase B: blind decoder training (the only gradient stage)**

initialize $\theta_d$; AdamW on decoder parameters only (backbone and libraries are constants)

**for** epoch $= 1, \ldots, 80$ **do**

  **for** batch (category-balanced), $x$ of class $c$, $(s, \ell) \in \mathcal{S} \times \{1,2,3,4\}$ **do**

    $M_{s,\ell} \leftarrow$ checkerboard mask ($s{\times}s$ blocks; each token masked once per scale) ▷ 8 forwards

    clean pass gives $z(u)$; $h \leftarrow \text{ViT}_\theta(P(x) \odot (1{-}M_{s,\ell}) + e_{\text{mask}} \odot M_{s,\ell} + p)$ ▷ Eq. 4: no $z(u)$ in $h$

    $q_u = \text{pe}(u) + e_s + h(u), \quad H \leftarrow \text{Dec}_{\theta_d}(\{q_u\}, h)$ ▷ Eq. 5

    $\ell_k(u) = s_\theta \left\langle \widehat{W_q H(u)},\, \widehat{W_k a_k} \right\rangle$ over $A_c \setminus \text{own}(x)$ ▷ Eq. 6: training-time blinding

    $\pi_k(u) \leftarrow (1{-}\varepsilon)\,\text{softmax}_k\, \ell(u) + \varepsilon / K_c$ ▷ Eq. 7

    $E^{(s,\ell)}(u) \leftarrow -\tau \log \sum_k \pi_k(u)\, \exp\big((\langle z(u),\, a_k \rangle - 1)/\tau\big)$ ▷ vMF mixture free energy

    accumulate $\mathcal{L} \leftarrow$ mean of $E^{(s,\ell)}(u)$ over forwards and target positions ▷ Eq. 8

  **end for**

  backpropagate to $\theta_d$ only; clip gradient norm

**end for**

select best epoch on $\mathcal{D}_{\text{val}}$ (validation energy); freeze $\theta_d$

**Phase C: leakage-safe calibration (no gradients)**

witness fields for all $x \in \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{val}}$: $d_x(u) = \max\big(0,\, 1 - \max_{r \in R_{c(x)}} \langle z_x(u),\, r \rangle\big)$ ▷ Eq. 10

route nulls on $\mathcal{D}_{\text{cal}}$: pool $\{d_x(u)\}$ over all $P$ positions per route $r \to (\hat{\mu}_r, \hat{\sigma}_r)$ ▷ Eq. 13

image-score null on $\mathcal{D}_{\text{cal}}$: $s(x_i) = \text{CVaR}_{0.01}[d_{x_i}]$; per-route sorted $\{s(x_i)\}_{i \le N_r}$ ▷ Eq. 11

$s_{\text{val}} \leftarrow \text{Std}\big(\{\tilde{z_{\text{G}}}(u) : x \in \mathcal{D}_{\text{val}},\, u\}\big), \quad \tilde{z_{\text{G}}} = (d_x - \hat{\mu}_{c(x)})/\hat{\sigma}_{c(x)}$ ▷ Eq. 14: scale correction

mode moments on $\mathcal{D}_{\text{val}}$: $m(u) = \arg\max_j \langle z(u),\, \rho_j \rangle$, shrunk $(\mu_{c,m}, \sigma_{c,m})$ ▷ Eq. 9

freeze and hash-lock all statistics

## F PER-CATEGORY RESULTS

Tables 8 and 9 report per-category pixel AUROC, AP, and AUPRO (mean over three seeds) of the contextual field alone and of the full TRIAD joint field. The per-category AUROC gains are visualized in Figure 2 of the main text.

## G QUALITATIVE ANOMALY MAPS

Figures 6–9 visualize the deployed TRIAD model on every category of both datasets. For each category we show the normal test image with the lowest image-level witness score and the anomalous test image with the highest image-level witness score, deterministically selected without manual curation, together with the ground-truth mask and the overlaid pixel anomaly map. The map is the calibrated joint field $J$ of Eq. 15, rendered through the frozen pixel path of bilinear upsampling to

**Algorithm 2** TRIAD inference: category routing, tri-evidence computation, and calibrated outputs (a single frozen sweep; no test-time fitting)

**Require:** test image $x$; all frozen artifacts of Algorithm 1; decision level $\alpha'{=}0.05$
**Ensure:** pixel anomaly map $A(x) \in [0,1]^{448\times448}$; image score $s(x)$; $p$-value $\hat{p}(x)$; binary decision

**Step 0: category routing (deterministic, frozen)**

whole-view pass at $448^2 \to$ last-layer tokens $h_{12}(t)$; semantic descriptor $s(x)$ of Eq. 2

$c(x) \leftarrow \arg\max_k \langle s(x),\, \nu_k \rangle$ ▷ Eq. 3: selects $A_{c(x)}$, $\{\rho_j\}$, $R_{c(x)}$, and the route nulls

**Step 1: contextual energy branch**

compose $896^2$ canvas from four native $448^2$ views ($64{\times}64$ tokens); clean pass gives $z(u)$

**for** $(s,\ell) \in \{2,4\} \times \{1,2,3,4\}$ **do**

$h \leftarrow \mathrm{ViT}_\theta(P(x) \odot (1{-}M_{s,\ell}) + e_{\mathrm{mask}} \odot M_{s,\ell} + p)$ ▷ Eq. 4: $z(u)$ never enters $h$

$q_u = \mathrm{pe}(u) + e_s + h(u)$; $H \leftarrow \mathrm{Dec}_{\theta_d}(\{q_u\}, h)$; logits over $A_{c(x)}$ ▷ Eqs. 5–6

$\pi_k(u) \leftarrow (1{-}\varepsilon)\,\mathrm{softmax}_k\, \ell(u) + \varepsilon / K_{c(x)}$ ▷ Eq. 7

$E^{(s,\ell)}(u) \leftarrow -\tau \log \sum_k \pi_k(u)\, \exp\big(\langle z(u),\, a_k \rangle - 1)/\tau\big)$ ▷ vMF mixture free energy

**end for**

$E(u) \leftarrow \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} E^{(s)}(u)$; $m(u) \leftarrow \arg\max_j \langle z(u),\, \rho_j \rangle$ ▷ Eq. 8

$z_{\mathrm{V}}(u) \leftarrow \big(E(u) - \mu_{c(x),m(u)}\big)/\sigma_{c(x),m(u)}$ ▷ Eq. 9; $z_{\mathrm{V}} \in \mathbb{R}^{64\times64}$

**Step 2: witness branch**

whole-view pass $\to P{=}1024$ query descriptors $z_x(u)$, $u = 1,\ldots,P$

$d_x(u) \leftarrow \max\big(0,\, 1 - \max_{r\in R_{c(x)}} \langle z_x(u),\, r \rangle\big)$ ▷ Eq. 10; $d_x \in [0,1]^{32\times32}$

$s(x) \leftarrow \mathrm{CVaR}_{0.01}[d_x] = \frac{1}{11}\sum_{i=1}^{11} d_{(i)}(x)$, $d_{(1)} \geq d_{(2)} \geq \cdots$ ▷ Eq. 11: tail mean

$\hat{p}(x) \leftarrow \frac{1+\#\{i:\, s(x_i)\geq s(x)\}}{N_{c(x)}+1}$ ▷ Eq. 12: add-one rule; finite-sample valid

**Step 3: calibrated joint branch**

$\tilde{z}_{\mathrm{G}}(u) \leftarrow (d_x(u) - \hat{\mu}_{c(x)})/\hat{\sigma}_{c(x)}$; $z_{\mathrm{G}}(u) \leftarrow \tilde{z}_{\mathrm{G}}(u)/s_{\mathrm{val}}$; upsample $32 \to 64$ ▷ Eqs. 13–14

$J(u) \leftarrow \frac{1}{2}\big(z_{\mathrm{V}}(u) + z_{\mathrm{G}}(u)\big)$ ▷ Eq. 15: LR-optimal equal weights

$A(x) \leftarrow \mathrm{Gauss}_{\sigma=4}\big(\mathrm{bilinear}_{448}(J)\big)$ ▷ frozen pixel path

**return** $A(x)$, $s(x)$, $\hat{p}(x)$; anomaly iff $\hat{p}(x) \leq \alpha'$ ▷ detection bit-identical to the witness branch

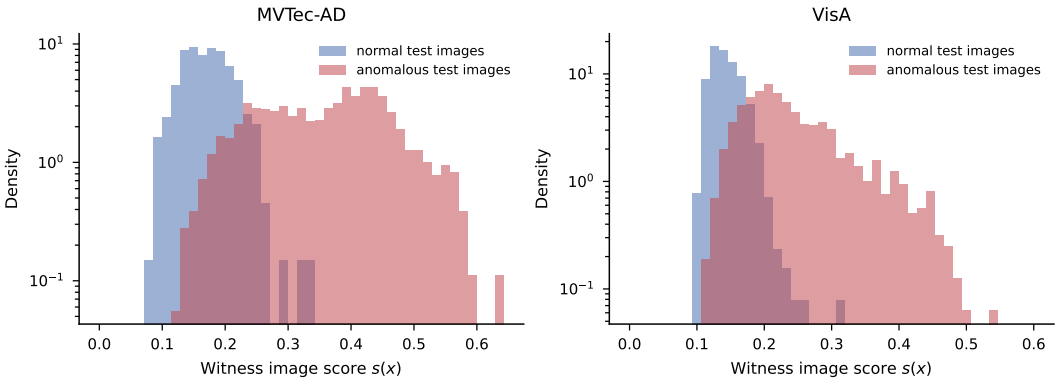


Figure 4: Witness image-score distributions for normal and anomalous test images, pooled over all categories of each dataset.

448 and Gaussian smoothing with $\sigma{=}4$. For display only, each map is min–max normalized per image and colored with the perceptually uniform `inferno` colormap; normal rows have empty ground-truth masks. A selection manifest with the exact image paths, scores, and $p$-values per category is released alongside the figures for auditability. The overlays reproduce the quantitative picture: responses concentrate on the defective regions, including small and low-contrast defects, while normal images show no salient activation.

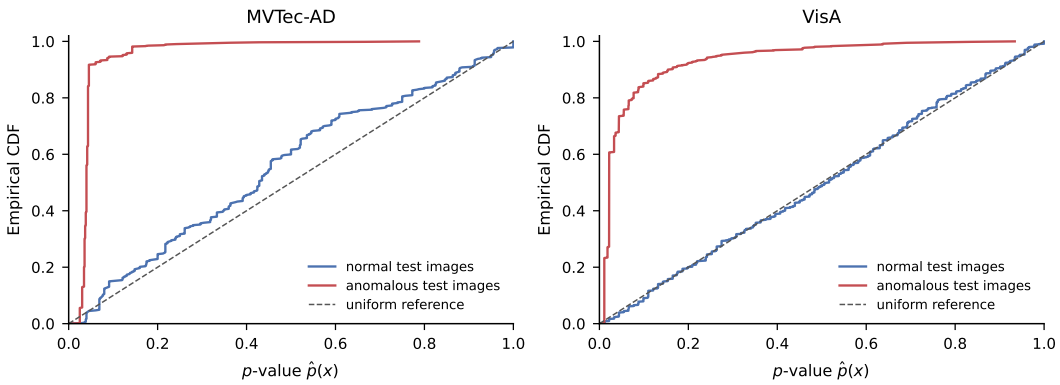


Figure 5: Empirical CDF of the add-one $p$-values. Normal test images follow the uniform reference, confirming the finite-sample validity of Eq. 12, while anomalous images concentrate near zero.

Table 8: Per-category results on MVTec-AD (pixel level, mean over three seeds).

| | AUROC | | AP | | AUPRO | |
|---|---|---|---|---|---|---|
| Category | contextual | TRIAD | contextual | TRIAD | contextual | TRIAD |
| bottle | 98.83 | **99.29** | 84.79 | **88.33** | 96.63 | **97.89** |
| cable | 97.43 | **98.76** | 72.68 | **77.20** | 92.96 | **95.95** |
| capsule | 97.23 | **98.90** | 55.48 | **60.73** | 95.30 | **96.85** |
| carpet | 98.04 | **99.52** | 64.72 | **70.02** | 96.82 | **98.71** |
| grid | 99.55 | **99.63** | **61.11** | 60.78 | 97.16 | **97.76** |
| hazelnut | 99.11 | **99.62** | 77.42 | **81.92** | 92.96 | **97.20** |
| leather | **99.47** | **99.47** | **47.89** | 47.53 | **98.61** | **98.64** |
| metal_nut | 97.04 | **98.06** | 80.08 | **83.65** | 93.25 | **96.32** |
| pill | 92.50 | **95.97** | 57.52 | **65.16** | 95.50 | **97.47** |
| screw | 96.35 | **99.14** | 24.51 | **56.19** | 88.87 | **96.13** |
| tile | 97.44 | **97.73** | **71.65** | 69.18 | 93.46 | **93.87** |
| toothbrush | 98.74 | **99.30** | 47.26 | **55.00** | 95.18 | **97.23** |
| transistor | 96.29 | **97.60** | 71.86 | **74.86** | 91.28 | **94.12** |
| wood | 95.47 | **96.40** | 69.75 | **72.39** | 95.69 | **96.40** |
| zipper | 97.73 | **97.97** | **63.68** | 60.75 | 94.04 | **94.64** |
| **macro** | 97.42 | **98.49** | 63.36 | **68.25** | 94.51 | **96.61** |

# H ADDITIONAL ABLATIONS, FALSIFIED DIRECTIONS.

## H.1 ABLATION SUMMARY FIGURE

Figure 10 summarizes the three preregistered ablation suites of the main text.

## H.2 ENTROPY CONDITIONING (FALSIFIED)

We preregistered and tested an alternative third evidence: re-weighting the contextual field by the entropy of the blind atom posterior, hypothesizing that high-entropy, ambiguous positions carry more background false positives. The diagnostic falsified the hypothesis before deployment. High-entropy positions enrich true anomalies, with a true-positive share of $0.53$ to $0.91$, more than background errors, with a share of $0.50$ to $0.77$. The posterior entropy is itself anomaly signal, and conditioning on it removes signal instead of noise, with a macro pixel AUROC drop of $0.40$. We report this failed direction as part of the design-space evidence.

## H.3 IMAGE-LEVEL FUSION (FALSIFIED)

Fusing additional image-level evidence channels into the witness decision was ablated in an earlier cycle and degraded MVTec-AD image AUROC, since a saturated $99.29$ leaves no headroom. This is why TRIAD freezes the image decision rather than fusing at that level.

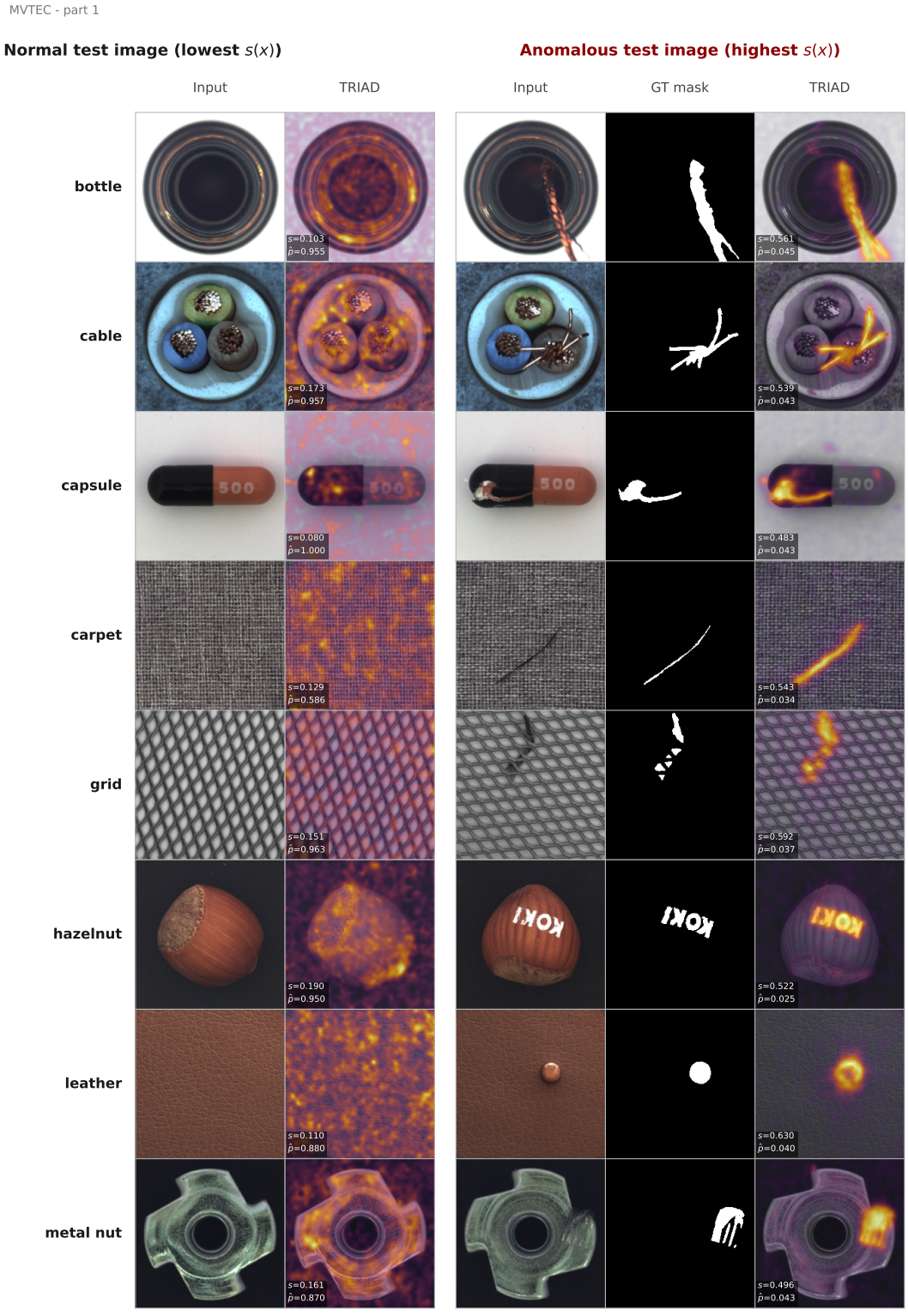


Figure 6: TRIAD qualitative results on MVTec-AD (categories 1–8): input, ground-truth mask, and the overlaid full-model pixel anomaly map.

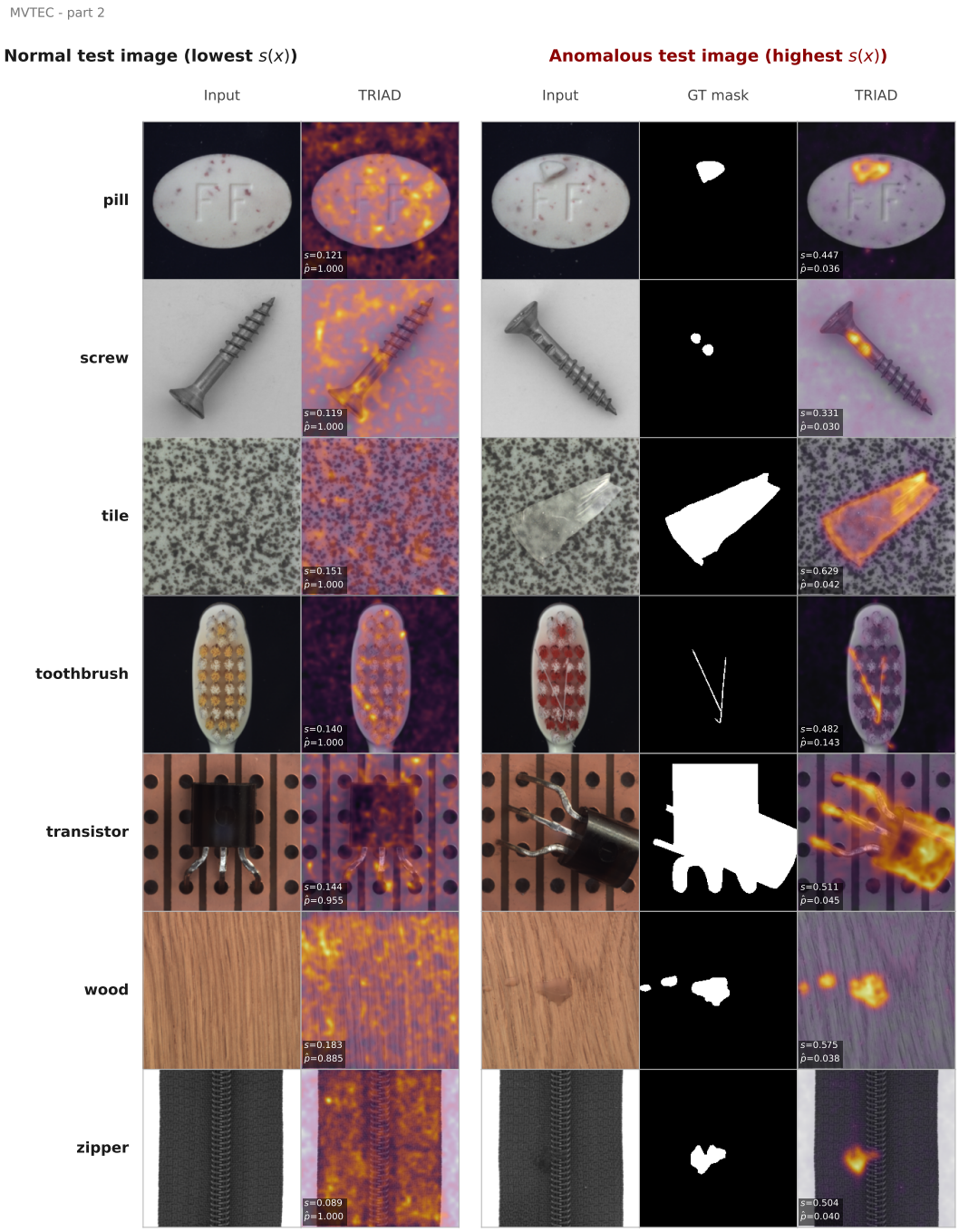


Figure 7: TRIAD qualitative results on MVTec-AD (categories 9–15).

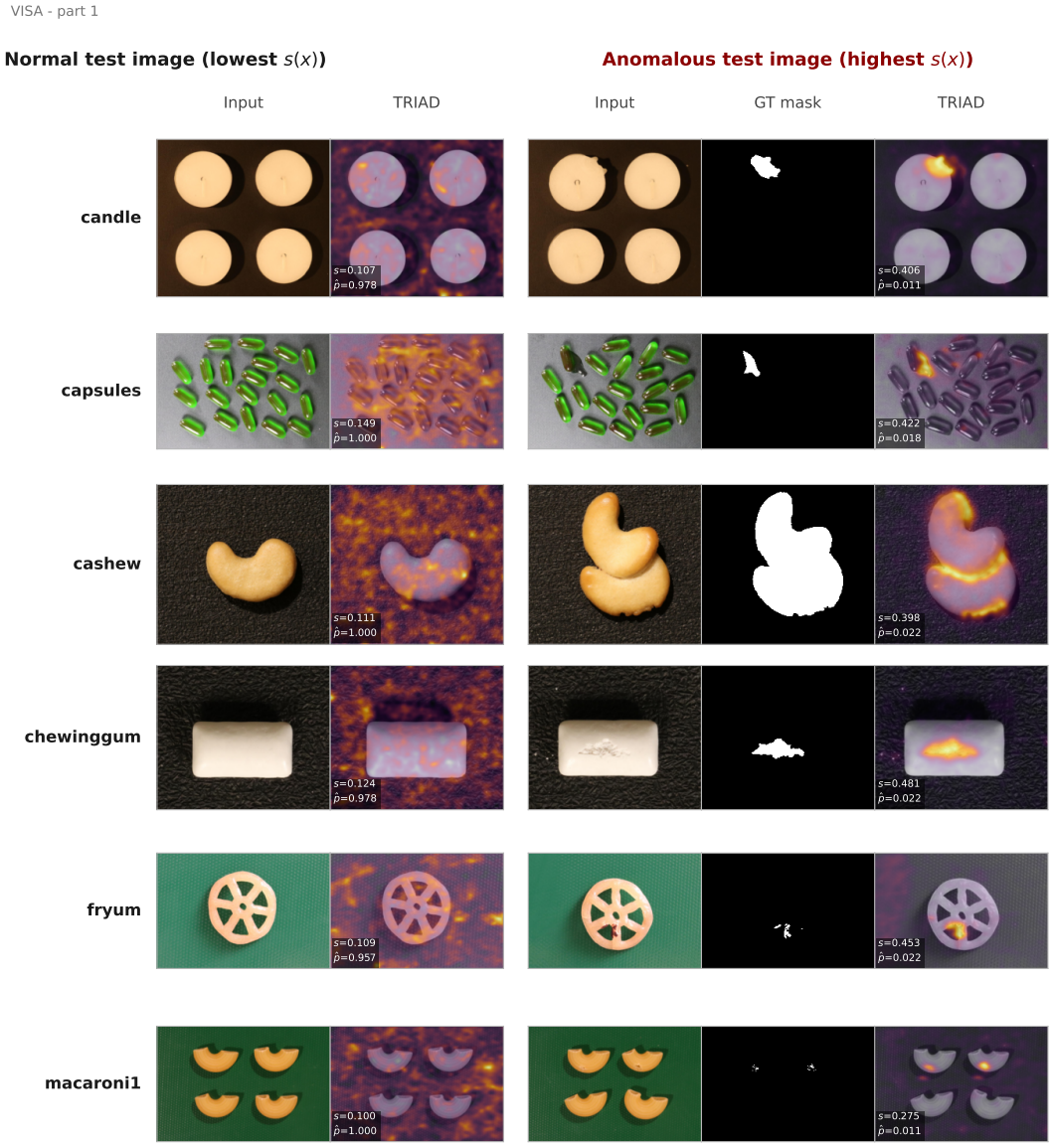


Figure 8: TRIAD qualitative results on VisA (categories 1–6).

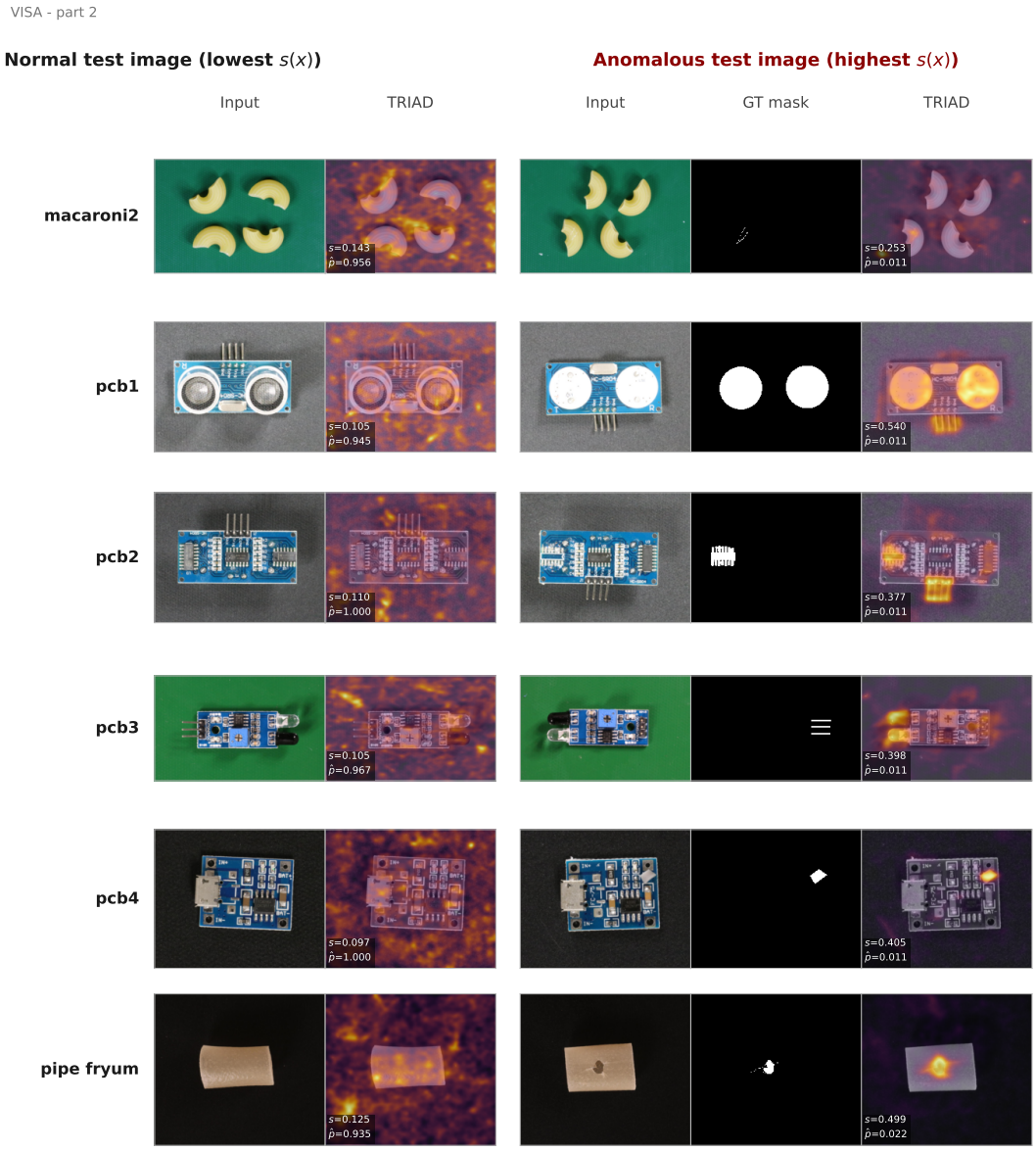


Figure 9: TRIAD qualitative results on VisA (categories 7–12).

Table 9: Per-category results on VisA (pixel level, mean over three seeds).

| | AUROC | | AP | | AUPRO | |
|---|---|---|---|---|---|---|
| Category | contextual | TRIAD | contextual | TRIAD | contextual | TRIAD |
| candle | 99.36 | **99.62** | 41.18 | **44.27** | 96.65 | **97.75** |
| capsules | 96.85 | **98.88** | 39.08 | **54.12** | 92.20 | **97.84** |
| cashew | 96.41 | **99.32** | 58.07 | **69.89** | 92.51 | **96.72** |
| chewinggum | **99.54** | **99.54** | 77.24 | **78.87** | 94.45 | **95.27** |
| fryum | 88.93 | **95.98** | 31.47 | **41.47** | 82.41 | **92.15** |
| macaroni1 | 99.08 | **99.17** | 16.23 | **19.06** | **93.41** | 92.66 |
| macaroni2 | 97.37 | **99.40** | 7.97 | **17.74** | 91.67 | **96.06** |
| pcb1 | 99.38 | **99.74** | 76.12 | **78.27** | 90.25 | **96.41** |
| pcb2 | 98.42 | **99.10** | 21.87 | **23.35** | 88.04 | **92.92** |
| pcb3 | 98.76 | **99.17** | **28.50** | 25.77 | 82.82 | **88.20** |
| pcb4 | 96.04 | **98.49** | 38.62 | **49.11** | 84.75 | **91.13** |
| pipe_fryum | 98.00 | **99.21** | 54.03 | **57.42** | 95.54 | **97.70** |
| **macro** | 97.34 | **98.97** | 40.86 | **46.61** | 90.39 | **94.57** |

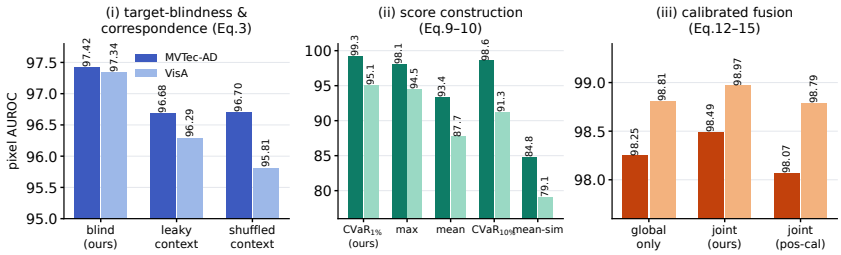


Figure 10: Ablation summary over three seeds. (i) Removing target blindness (leaky) or context correspondence (shuffled) degrades localization on both datasets. (ii) The $\text{CVaR}_{1\%}$ tail readout over a nearest-neighbour witness field dominates every alternative score construction. (iii) The calibrated fusion beats witness-only aggregation and the per-position calibration ablation.

### H.4 Single-scale masking schedules

Restricting the schedule to scale 2 or to scale 4 alone changes MVTec-AD pixel AUROC to 97.41 and 97.39, against 97.42 for the deployed multi-scale schedule. The two single-scale variants trade small advantages on different categories, which the multi-scale average removes.