How Much of Anomaly Segmentation Is Background? A Published 97% Pixel AUROC Can Certify Chance Inside the Object

Anonymous authors
Paper under double-blind review

Abstract

A published 97% pixel AUROC on liver CT certifies chance inside the organ. Anomaly segmentation, medical and industrial alike, is scored almost universally by that number over the whole image, and we show what it measures. On liver CT a Gaussian blur of the input, with no model, no training and no support set, scores 94.8% and a binary foreground mask 96.9%, within three points of published detectors; inside the organ they score 14.0% and chance. The reason is an identity: whole-image AUROC equals $wA_{\mathcal{B}} + (1-w)A_{\mathcal{F}}$, weighted by the fraction w of normal pixels that are background, so a reported score certifies only $A_{\mathcal{F}} \geq (A-w)/(1-w)$; on liver CT w=0.94, and the claim is about w: on breast ultrasound (w=0.15) the same 97% certifies $A_{\mathcal{F}} \geq 96.5\%$. Class imbalance is not the mechanism, and balancing the classes leaves the number unchanged; the partition is a reporting choice that leaves A untouched, and across 6 partition rules the liver-CT guarantee moves by 2.6 points. The clearest evidence is a single detector: masking the background out of a nearest-neighbour memory takes its whole-image score on liver CT from 97.1 to 4.5 while its foreground score barely moves, 64.7 to 63.4 — same weights, same images, only the negative class changed. Measuring every term on 53,260 images across 8 dataset families spanning medicine, industry, remote sensing and driving scenes, the inflation follows w wherever a predictor can separate the background and vanishes where it cannot. Re-running training-free detectors on the six medical benchmarks under both metrics, the ranking they induce changes wherever the frame is empty (liver CT: Kendall τ=0.12, 40 of 91 pairs flip) and holds where it is full (breast ultrasound: τ=0.98). We recommend reporting $A_{\mathcal{F}}$ beside A with the partition stated, no new metric but the same statistic on a stated set of normals, release the audit as a one-file tool, and report a training-free reference baseline both ways.

1 Introduction

Anomaly segmentation asks a model to mark the pixels of an image that are not normal: a lesion in a scan, a scratch on a part. Because the positive pixels are rare, the field scores the map with a threshold-free, prevalence-free statistic, the area under the ROC curve over every pixel of every test image, and reports it as *pixel AUROC*. On the medical benchmarks that decide few-shot anomaly detection (Bao et al., 2024; Huang et al., 2024) the number sits in the high nineties for most methods, as on the industrial ones (Bergmann et al., 2019; Zou et al., 2022).

Fig. 1 shows what that number can measure: on liver CT a Gaussian blur of the scan reaches 94.8% pixel AUROC and a binary foreground mask 96.9%, within three points of the best published method; against the organ's own pixels the blur scores 14.0% and the mask, constant inside the organ, scores chance. The same predictor on an industrial object on a tray behaves the same way (Fig. 1, bottom); nothing in what follows is medical.

That the metric flatters segmentation has been remarked on before (Meissen et al., 2021; Bertoldo et al., 2024; Cai et al., 2025), and the explanation usually offered is class imbalance. It is the wrong

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.

explanation, and it prescribes the wrong remedy: balancing the classes leaves the number unchanged (Section 3). AUROC is the probability that a random anomalous pixel outscores a random normal one; it does not depend on how many of each there are, only on *which* normal pixels there are. Partition the normal pixels into background $\mathcal{B}$ and foreground $\mathcal{F}$ and let w be the fraction that are background. Then, exactly,

$$\text{AUROC} \;=\; w\,A_{\mathcal{B}} \;+\; (1-w)\,A_{\mathcal{F}}, \tag{1}$$

with $A_{\mathcal{B}}$ the AUROC of anomalous against background pixels and $A_{\mathcal{F}}$ against foreground pixels (Section 2). The proof is one line; the consequences are not. Because a black background is separable by almost any predictor, $A_{\mathcal{B}} \approx 1$, and the reported number contains a floor of w before any discrimination inside the tissue has happened: 93.9% on liver CT. Since the metric moves by $1-w$ per point of $A_{\mathcal{F}}$, 16.5 points of foreground discrimination there are worth one point in the table. And since $A_{\mathcal{B}} \le 1$, a published number certifies only $A_{\mathcal{F}} \ge (A-w)/(1-w)$: on liver CT a 97% certifies chance inside the organ and a 99.8% certifies $A_{\mathcal{F}} \ge 96.7\%$. The statement is about w: on breast ultrasound, where w=0.15, the same 97% certifies $A_{\mathcal{F}} \ge 96.5\%$. The partition is a reporting choice that leaves A untouched; across all 6 partition rules we test, the liver-CT guarantee moves by 2.6 points.

The identity says what *can* happen; the audit measures what *does*. We compute every term of Eq. 1 for the same two model-free predictors on $53{,}260$ images from 8 dataset families: every public two-dimensional benchmark with pixel-level anomaly masks we could obtain, across medicine, industry, remote-sensing change detection and driving scenes (Section 4). The residual of the identity is at floating-point precision on every row. Wherever the estimator finds a background that the predictor separates, on scans, on objects photographed on a tray, on uniform textures, the inflation follows w (Spearman ρ=0.89 over all 64 categories) and reaches 80.8 points; where the background is not separable to the predictor, as for a difference map in change detection, the inflation stays near zero whatever w is, exactly as $w(A_{\mathcal{B}} - A_{\mathcal{F}})$ says it must.

An audit matters if it changes conclusions. Section 5 runs 12 training-free detectors and the two model-free predictors through one code path on the six medical benchmarks and ranks them under A and under $A_{\mathcal{F}}$: on liver CT 40 of 91 detector pairs change order, a detector at 95% whole-image AUROC sits at chance inside the organ, and under A the model-free blur outranks real detectors. The recommendation is short: report $A_{\mathcal{F}}$ beside A, state the partition, and compare methods on $A_{\mathcal{F}}$.

Contributions. (i) The exact decomposition of pixel AUROC into background and foreground terms and what it yields that restricting a score to a region does not: a floor, an exchange rate, a guarantee that re-reads every published number without re-running anything, and the mechanism, background separability rather than class imbalance (Section 2). (ii) A model-free audit of $53{,}260$ images across 8 dataset families, industrial and non-AD included (Sections 3–4). (iii) A re-ranking of 12 detectors showing that the foreground term changes which method wins (Section 5). (iv) A reporting convention, a tool, and a reference baseline reported under it (Section 6).

2 PIXEL AUROC DECOMPOSES EXACTLY

Let $\mathcal{P}$ be the set of anomalous pixels in a test set and $\mathcal{N}$ the set of normal pixels, pooled over images as the field pools them. A score map s assigns every pixel a real number. The empirical AUROC is the fraction of (anomalous, normal) pairs the score orders correctly, ties counting one half:

$$A(\mathcal{P},\mathcal{N}) \;=\; \frac{1}{|\mathcal{P}|\,|\mathcal{N}|}\sum_{p\in\mathcal{P}}\sum_{n\in\mathcal{N}}\Big[\mathbb{1}\{s(p)>s(n)\} + \tfrac{1}{2}\mathbb{1}\{s(p)=s(n)\}\Big], \tag{2}$$

the Mann–Whitney form that every AUROC implementation computes.

Proposition 1. *Let $\mathcal{N} = \mathcal{B} \sqcup \mathcal{F}$ be any partition of the normal pixels and $w = |\mathcal{B}|/|\mathcal{N}|$. Then for every score map, $A(\mathcal{P},\mathcal{N}) = w\,A(\mathcal{P},\mathcal{B}) + (1-w)\,A(\mathcal{P},\mathcal{F})$.*

Proof. Split the inner sum of Eq. 2 over $n \in \mathcal{B}$ and $n \in \mathcal{F}$; the parts are $|\mathcal{B}|\,A(\mathcal{P},\mathcal{B})$ and $|\mathcal{F}|\,A(\mathcal{P},\mathcal{F})$, and $|\mathcal{B}|/|\mathcal{N}| = w$. □

Nothing is assumed about the score, the images, the partition or the domain; the population version and the version on any pixel subsample hold by the same argument. We write $A_{\mathcal{B}} = A(\mathcal{P},\mathcal{B})$ and

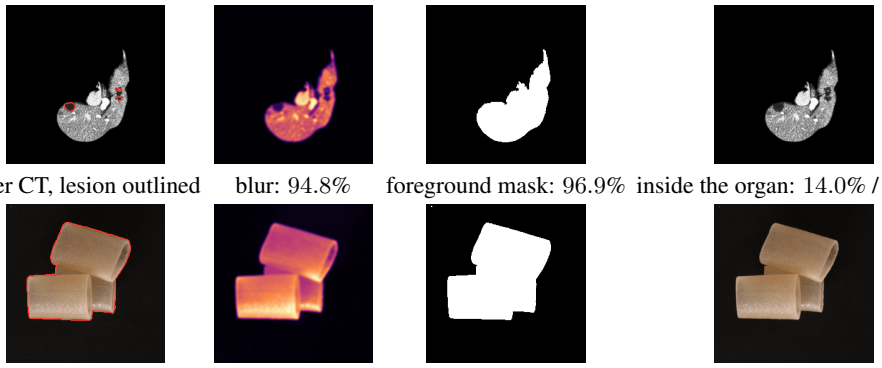


Figure 1: What whole-image pixel AUROC rewards. Top: a Gaussian blur of a liver CT scan and a binary foreground mask, neither containing a model, score 94.8% and 96.9% pixel AUROC on the benchmark, within three points of published detectors; against the organ's own pixels the blur scores 14.0% and the mask exactly chance. Bottom: the same predictor on a VisA object on a tray (w=0.89): 94.4% over the frame, 50.4% inside (mask $A_{\mathcal{B}}$=99.8). The whole-image number is not localisation; it is how small the object is in an empty frame.

$A_{\mathcal{F}} = A(\mathcal{P}, \mathcal{F})$. The proposition is elementary, and restricting a score to a region of interest is old. What is new is the relation it fixes between the restricted and the published number: on the benchmarks the field uses, one particular partition has $A_{\mathcal{B}}$ pinned near 1 for every predictor anyone would submit, which turns an algebraic identity into a statement about what a reported number can and cannot certify, and names the mechanism, background separability, that the usual explanation, class imbalance, gets wrong.

A floor. If a predictor ranks every anomalous pixel above every background pixel, $A_{\mathcal{B}} = 1$ and $A = w + (1 - w)A_{\mathcal{F}} \geq w$. A predictor constant on the foreground (a segmentation of the organ) has $A_{\mathcal{F}} = \frac{1}{2}$ and scores exactly $w + (1 - w)/2$; on liver CT that is 96.9%, which is what the mask in Table 1 obtains.

An exchange rate. $\partial A / \partial A_{\mathcal{F}} = 1 - w$. On liver CT one point of the reported metric costs $1/(1 - w) = 16.5$ points of discrimination inside the organ; on breast ultrasound (w=0.15) it costs 1.2. The same improvement inside the tissue is worth fourteen times less in the reported number on one dataset than on the other.

A guarantee. Since $A_{\mathcal{B}} \leq 1$, a reported A implies $A_{\mathcal{F}} \geq (A - w)/(1 - w)$. This is the only statement about foreground discrimination a whole-image number certifies. Table 14 evaluates it for published numbers.

Class imbalance is not the mechanism. Eq. 2 is an average over pairs, so subsampling either class uniformly leaves it unchanged in expectation, however extreme $|\mathcal{N}|/|\mathcal{P}|$. What changes it is the composition of $\mathcal{N}$: removing the background pixels takes w to zero and the metric from A to $A_{\mathcal{F}}$. Both statements are verified on data in Section 3.

Means over datasets. A benchmark mean of whole-image AUROC over datasets with different w averages $w_i A_{\mathcal{B}}^{(i)} + (1 - w_i) A_{\mathcal{F}}^{(i)}$, dominated by the datasets with the largest w, where foreground discrimination counts for least.

What "foreground" is. The identity holds for any partition, so the partition is a reporting choice and we state ours. One estimator is used for every image in this paper: the background is the region connected to the frame edge whose intensity is within 0.06 of the border's median, the foreground is everything else. It uses no model, label or learned threshold, and is the same rule for the black outside a scan and the tray under an object. Because $A_{\mathcal{F}}$ scores *all* anomalous pixels against the foreground normals, anomalous pixels the estimator puts in the background do not break the identity; on the six medical datasets it does so for at most 11.8% of lesion pixels. Section 3 sweeps the tolerance and an alternative intensity-threshold rule (Table 6): the terms move, the conclusions do not, and the one dataset where the rule matters is named. For remote-sensing change detection, where nothing is connected to the frame edge, Section 4 uses a different model-free partition and says so.

Table 1: Every term of the identity on the six medical mask datasets, for two predictors that contain no model, with the foreground estimated by the border rule (no model, no label, no learned threshold), on the full test splits. w is the fraction of normal pixels that are background; $A_{\mathcal{B}}$ and $A_{\mathcal{F}}$ are the AUROC of anomalous pixels against background and against foreground normals; the whole-image column is what the field reports and equals $wA_{\mathcal{B}} + (1-w)A_{\mathcal{F}}$ on every row to within 10^{-10}. The inflation is the whole-image score minus $A_{\mathcal{F}}$, i.e. $w(A_{\mathcal{B}} - A_{\mathcal{F}})$; the last column is the exchange rate $1/(1-w)$. Spearman ρ=0.83 (p=0.06) between w and the blur's inflation. Rows are ordered by w.

	fg.		Gaussian blur of the input				binary foreground mask				
Dataset	% frame	w	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl.	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl.	$\frac{1}{1-w}$
Liver CT	6.4	0.939	100.0	14.0	94.8	80.8	99.9	49.9	96.9	47.0	16.5
Retina RESC	25.3	0.781	97.2	42.0	85.1	43.0	94.1	44.1	83.1	39.0	4.6
Brain MRI	29.0	0.718	100.0	86.0	96.1	10.1	100.0	50.0	85.9	35.9	3.5
Head CT	78.3	0.226	99.7	73.0	79.0	6.0	100.0	50.0	61.3	11.3	1.3
Brain tumour	84.4	0.167	99.6	79.6	82.9	3.3	100.0	50.0	58.3	8.3	1.2
Breast US	84.9	0.154	89.5	39.8	47.5	7.7	94.2	44.2	51.9	7.7	1.2

3 THE AUDIT ON MEDICAL BENCHMARKS

We measure every term of Eq. 1 on the full test splits of the six medical benchmarks with lesion masks (brain MRI, liver CT, retinal OCT, head CT, a second brain MRI source, breast ultrasound; Appendix E) for two predictors that contain no model: the smoothed distance of each pixel's intensity from the frame border's intensity, which on a black-background scan is simply a Gaussian blur of the scan, and the estimated foreground itself as a binary mask. Both are "find whatever is not background"; neither has seen a lesion. AUROC is computed exactly from score histograms over every pixel of every image, not from a sample. Table 1 gives w, $A_{\mathcal{B}}$, $A_{\mathcal{F}}$, the whole-image score and the inflation $A - A_{\mathcal{F}} = w(A_{\mathcal{B}} - A_{\mathcal{F}})$; the residual of the identity is below 10^{-10} on every row.

What the table says. On liver CT the organ is 6.4% of the frame, w=0.94, and the blur has $A_{\mathcal{B}}$=100.0 and $A_{\mathcal{F}}$=14.0: perfect against the background, far worse than chance inside the organ (brighter parenchyma is less often lesion, so the blur ranks lesions low), 94.8% over the whole frame. On breast ultrasound tissue fills the frame, w=0.15, and the same blur scores 47.5% whole-frame and 39.8% inside, both below chance. On every dataset but breast ultrasound $A_{\mathcal{B}} \geq 97.2$, so on scans the inflation is essentially $w(100 - A_{\mathcal{F}})$: largest where the frame is emptiest *and* the predictor is weakest inside the tissue, which on liver CT is both. The inflation runs from 80.8 points to 3.3 and follows w in rank across the six (ρ=0.83); with $A_{\mathcal{B}}$ pinned at 1 that ordering is close to forced, and we do not lean on it. What is not forced is that $A_{\mathcal{F}}$ of a blur sits below chance on liver CT and at 86.0 on brain MRI, i.e. that the foreground term of a model-free predictor is small, so that what the reported numbers have been measuring on these benchmarks is mostly w.

The imbalance control. Table 10 (Appendix B) scores the same blur with the normal pixels subsampled at random to equal the anomalous ones, removing an imbalance of up to 219:1 while preserving w: the whole-image score moves by at most 0.88 points. Removing the background normals instead, which preserves most of the imbalance and sets w to zero, moves liver CT from 94.8 to 14.4. The metric is inflated by what the negative class is made of, not by how large it is.

Reading published numbers. Fig. 2 applies the guarantee, with w from each full test split, to every pixel AUROC published on the three BMAD segmentation datasets at the standard budget. On liver CT a reported 97% certifies nothing beyond chance, 99% certifies $A_{\mathcal{F}} \geq 83.5\%$, and the best published few-shot number, 99.8% at k=8 (Huang et al., 2024) (99.7 at k=16, the column Fig. 2 plots), certifies $A_{\mathcal{F}} \geq 96.5\%$. On brain MRI the same method's 97.2% certifies only $A_{\mathcal{F}} \geq 90.1\%$, because w is smaller there and the number sits closer to the floor. Concurrent training-free work reports 98.9% on liver CT (Huo et al., 2026); the bare foreground mask reaches 96.9% and the published number certifies $A_{\mathcal{F}} \geq 82.2\%$. The bound is the only thing the reported number establishes, which is why Section 5 re-runs detectors rather than re-reading their numbers.

The partition, and other metrics. Sweeping the border tolerance over 0.03–0.10 and replacing the rule by an intensity threshold at 0.05–0.20 (Table 6, Appendix B) moves individual terms and leaves every statement above standing; the one exception is the intensity threshold, which puts 25%

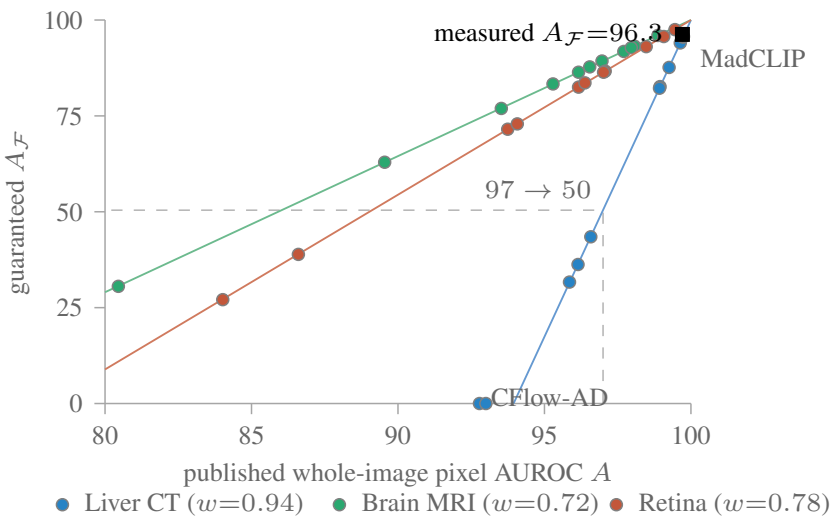


Figure 2: A published leaderboard, re-read. Every pixel AUROC published on the three BMAD segmentation datasets at the standard 16-shot budget (11 methods, 2022–2026, as tabulated by MadCLIP's Table 1 and HiMatch-AD's Table 2), against what it certifies inside the foreground, $(A - w)/(1 - w)$ with each dataset's full-split w. The points lie on three lines of slope $1/(1 - w)$. On liver CT the slope is 17: a leaderboard running from 92.8 to 99.7 certifies anything from 0 to 96 inside the organ, and the reported column cannot tell which. The dashed corner is the abstract's example. The filled square is the one point we measured rather than read: MVFA's released checkpoint decomposed on our pixel sample (Section 5), whose A=99.7 certifies 94.7 and whose $A_{\mathcal{F}}$ is 96.3, the arrow's length above its bound. Every point, its source and its bound: Table 15 (Appendix D).

of breast-ultrasound lesion pixels below it and is why we use the border rule. On liver CT all 6 rules agree: w within 0.939–0.942, the blur's $A_{\mathcal{F}}$ within 10.9–15.1. The background does not reach every metric the same way (Table 8). Per-image pixel AUROC obeys Eq. 1 image by image and moves with it: on liver CT, 94.5 whole-frame to 15.2 inside. AP and F1-max barely move (liver CT: AP 3.4 whole-frame and 3.4 inside, F1-max 11.2 and 11.4; every dataset alike), because a separable background sits below every real operating point. This is not a case for AP instead of $A_{\mathcal{F}}$: AP does not decompose, so it is not auditable by the same argument, and it is tied to prevalence as a rank statistic is not. Liver CT's small lesions give AP 3.4 and brain tumour's large ones 26.6 under the same restriction, a gap prevalence alone would produce from two equally good detectors and which $A_{\mathcal{F}}$, still an AUROC, does not confound. AUPRO inherits the identity on its FPR axis and is inflated more, not less: on liver CT the blur's AUPRO@0.3 is 83.2 as reported and 2.1 against foreground normals, and the identity predicts the 83.2 from the foreground curve alone (Appendix B). We tested this directly on the 13 detectors of Table 18: ranking by pixel AP restricted the same way agrees with the $A_{\mathcal{F}}$-ranking at τ=0.79 on liver CT, and τ=0.54–0.87 across all six datasets, against A's own agreement with $A_{\mathcal{F}}$ of τ=0.12–0.98 on the same detectors. AP already sees through this failure mode; the identity is what explains why, and $A_{\mathcal{F}}$ is the version of the argument that also certifies published pixel-AUROC numbers.

4 Across domains

Six datasets, all medical, invites two objections: the sample is small, and the effect might be about medical images rather than about framing. Eq. 1 contains nothing medical, so we repeat the measurement on every public two-dimensional dataset with pixel-level anomaly masks we could obtain, under a criterion stated so that this is not convenience sampling: public, two-dimensional, per-pixel masks, at least one anomalous test image per category. The predictor is that of Section 3, the smoothed distance from the border intensity (on a light tray, the blur's negative). That yields 53,260 images in 8 families: the six medical benchmarks; 4 colonoscopy polyp sets (every image anomalous); the industrial benchmarks MVTec AD (Bergmann et al., 2019) (15 categories), VisA (Zou et al., 2022) (12), DTD-Synthetic (Cimpoi et al., 2014; Jeong et al., 2023) (12) and DAGM (10); 4 bi-temporal building change detection benchmarks; and the RoadAnomaly driving set. No detector

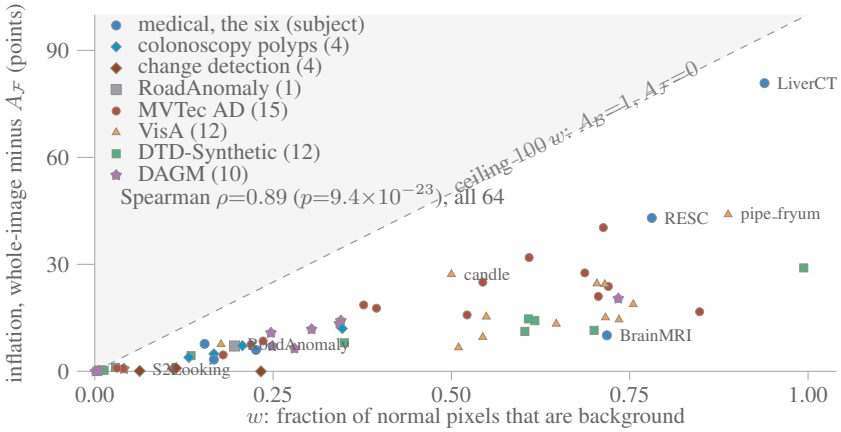


Figure 3: Inflation of whole-image pixel AUROC against the background fraction of the normal pixels, w, for every category of every dataset family audited, on full test splits. Every point is the same model-free predictor scored two ways; no detector appears in this figure. The shaded region is what Eq. 1 permits: the inflation is $w(A_{\mathcal{B}} - A_{\mathcal{F}})$ and cannot exceed $100\,w$. Categories whose background the predictor separates ($A_{\mathcal{B}}{\approx}1$) sit near the ceiling whatever their domain; change detection, where a difference map does not separate the vegetation background ($A_{\mathcal{B}}{\approx}0.6$), sits on the floor whatever its w.

is involved: every number is the same model-free predictor, and on every row the identity's residual is at floating-point precision (Appendix A).

Industrial benchmarks. On MVTec AD the model-free predictor reaches 91.1% on its best category and its inflation reaches 40.3 points on the objects, while on the textures (carpet, grid), where the rule finds almost no background ($w{<}0.05$), there is nothing to inflate. On VisA, where every object sits on a uniform tray, it reaches 94.8% on `macaroni2` ($w{=}0.74$, $A_{\mathcal{B}}{=}100.0$, $A_{\mathcal{F}}{=}80.2$) and 44.0 points of inflation on `pipe_fryum`. On the synthetic textures of DTD-Synthetic and DAGM the inflation follows w within each ($\rho{=}0.97$ and 0.95) and is near zero where w is. Within MVTec AD $\rho{=}0.77$ and within VisA, framed almost identically, $\rho{=}0.48$. Over all 64 categories, $\rho{=}0.89$ ($p{=}9.4{\times}10^{-23}$), and Fig. 3 shows every family against the ceiling $100\,w$. What the domains share is not medicine or industry but a background that a predictor with no model can separate.

Two domains outside anomaly detection. Dense evaluation by pixel AUROC is not confined to AD, so we audit two neighbours. In bi-temporal building change detection (LEVIR-CD, WHU-CD, DSIFN-CD, S2Looking; 21,280 test tiles) the model-free change map is the smoothed absolute difference of the histogram-matched pair, and the partition calls a pixel background if it is vegetation in *both* epochs by the excess-green index, where building change cannot occur. There w runs 0.06–0.23 but $A_{\mathcal{B}}$ only 56.7–78.9: seasonal vegetation differences look like change to a pixel-difference map, so the background is *not* free and the inflation stays at -0.6–0.9 points (Table 13, Appendix C). We tested a *trained* detector rather than assert one: BIT (Chen et al., 2021), the authors' released LEVIR-CD checkpoint, gives $A_{\mathcal{B}}{=}99.5$ and $A_{\mathcal{F}}{=}99.4$ on the same partition: near-ceiling, but *both* of them, so the inflation is 0.0 points because a strong detector's foreground discrimination leaves nothing for w to inflate, not because the partition failed (Appendix C) — the sharpest contrast in the paper with liver CT, where every detector's $A_{\mathcal{F}}$ is under 70. In driving-scene anomaly segmentation (RoadAnomaly (Lis et al., 2019)) the border rule finds the sky ($w{=}0.20$); the blur's terms are in Table 3. The standard detector of that literature, maximum-softmax over a Cityscapes-trained SegFormer (Xie et al., 2021), gives $A_{\mathcal{B}}{=}80.6$ on the same partition: the sky a blur separates at 97.3 is not separable by the segmenter's confidence, whose known failure mode is confidence on the anomaly itself, so its inflation is 2.4 points ($A_{\mathcal{F}}{=}68.4$; Table 5, Appendix A). On SegmentMeIfYouCan's 10+30 public labelled frames the pattern repeats (MSP inflation 0.8 and 0.2 points); on the obstacle track $A_{\mathcal{F}}{=}97.6$, so, as with BIT, there is nothing left to inflate.

Correlated is not predictable, so we tested the stronger claim. Holding out one category at a time, an isotonic regression of the inflation on w predicts the held-out inflation to 5.9 points mean absolute error, 40% below predicting the held-in mean (9.8), and misses the most inflated categories: useful for anticipating a number, not for correcting one, which needs w, the partition and

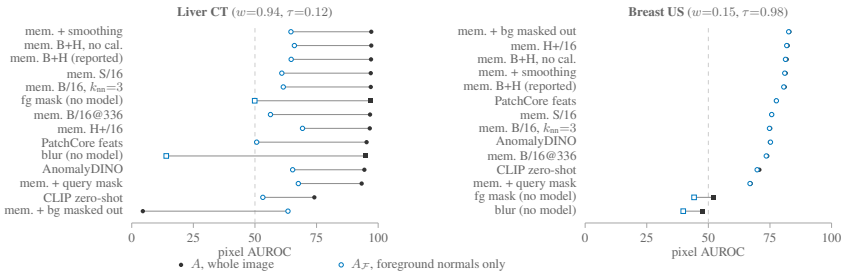


Figure 4: Every detector of this section on liver CT and on breast ultrasound, one row each, sorted by the reported metric: the black end of a row is its whole-image pixel AUROC A, the blue end its foreground term $A_{\mathcal{F}}$ on the same pixels. On liver CT, where 94% of the normal pixels are background, the reported ends of the patch memories stack into a column at 96.6–97.2 while the foreground ends spread from 14 to 69, and the rows cross; on breast ultrasound, with no background to order by, the rows are short and parallel. Squares: the two model-free predictors; dashed: chance. All six datasets, with ranks and pair flips: Table 18 (Appendix F).

$A_{\mathcal{B}}$ — hence report $A_{\mathcal{F}}$ rather than adjust A. A family audited only after every number here was fixed, dermoscopy (ISIC 2018, w=0.43), lands 1.1 points from the prediction (16.2 predicted, 17.3 measured; Appendix A).

5 Does the foreground term change the ranking?

The guarantee reads published numbers; a ranking needs score maps. We therefore run 12 detectors and the two model-free predictors through one evaluation code path on the six medical benchmarks and score each under A and under $A_{\mathcal{F}}$ on the same pixel sample (Fig. 4; every number, rank and flip is Table 18 in Appendix F). The detectors span the training-free design space within one framework (patch memories over frozen DINOv3 priors at three scales (Siméoni et al., 2025), the AnomalyDINO recipe (Damm et al., 2025), PatchCore's features (Roth et al., 2022) given the same k supports, the two-tap memory of Section 6 with and without calibration, with background masking, query masking and map smoothing), a zero-shot vision–language detector from a different family (the dense branch of VAND/APRIL-GAN over CLIP (Radford et al., 2021; Chen et al., 2023), no support set), and the two model-free predictors, ranked with the rest. Each memory uses k=16 unlabelled normals per dataset, two draws averaged; no arm sees an anomaly label; methods we could not run are read through Table 14.

The ranking changes where the frame is empty. Kendall's τ between the A-ranking and the $A_{\mathcal{F}}$-ranking runs from 0.12 on liver CT (0.12–0.19 over 12 replicates, Appendix F) to 0.98 on breast ultrasound, and 76 of 6×91 ordered pairs flip. Liver CT (w=0.94) is the case the identity predicts, and 40 of its 91 pairs change order. The foreground mask, which contains no model, ranks 6 of 14 under A at 96.9 and *last* under $A_{\mathcal{F}}$ at 49.9, chance by construction; PatchCore's features, 95.3 and 50.7, chance again behind a 95. The zero-shot vision–language detector scores 74.1 under A, *below* the model-free blur, and 53.2 under $A_{\mathcal{F}}$, 39 points *above* it: the order of a blur and a CLIP detector reverses with the metric. Masking the background out of the memory puts it last under A at 4.5 and 7 of 14 under $A_{\mathcal{F}}$ at 63.4. Where the frame is full (breast ultrasound, τ=0.98) the two columns agree; where it is empty, the whole-image column orders detectors by how cleanly they find the organ; the foreground column orders them by how well they find the lesion, and those are different orders.

What the foreground term rewards. The literature's additions show the mechanism directly. Masking background tokens out of the memory, AnomalyDINO's own recipe, takes the two-tap memory on liver CT from A=97.1 to A=4.5 while $A_{\mathcal{F}}$ moves from 64.7 to 63.4: with no background tokens to match, background query tokens score as maximally anomalous, $A_{\mathcal{B}}$ collapses, and the term that carried w=0.94 of the metric stops being fed (Appendix M.1). The same detector is therefore last under A and near the top under $A_{\mathcal{F}}$, which is the clearest statement of what the two columns measure. A method developer optimising A on these benchmarks is optimising the organ mask. Measured directly, every re-run arm with A in 93–97 overlaps the organ mask at Dice 53–81 and the masked-out memory (A=4.5) at 0.1, all at the same $A_{\mathcal{F}}$ (Table 19, Appendix F).

Table 2: Published methods through the identity: released checkpoints run with the authors' code (MVFA with its supports and zero-/few-shot fusion; AnomalyCLIP zero-shot, its MVTec-trained prompts), scored on the same pixel sample and partition as every training-free arm, on the three BMAD datasets MVFA's checkpoints cover. MVFA is in a different supervision class and leads under both metrics; the point is the size of each foreground term next to its whole-image number. Ranks in brackets over 16 rows; Kendall's τ and pair flips (of 120) in the last rows.

	Liver CT		Brain MRI		Retina	
Detector	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$
Gaussian blur of the input (no model)	94.8 (11)	14.0 (16)	96.1 (8)	86.0 (11)	85.1 (14)	42.0 (16)
foreground mask (no model)	96.9 (7)	49.9 (14)	85.9 (14)	50.0 (16)	83.1 (15)	44.1 (15)
CLIP zero-shot, prompts only (VAND/APRIL-GAN branch)	74.1 (15)	53.2 (12)	74.5 (15)	53.7 (15)	61.8 (16)	71.2 (14)
PatchCore features (WRN50-2), k supports	95.3 (10)	50.7 (13)	96.0 (9)	86.1 (10)	96.4 (9)	88.3 (9)
AnomalyDINO recipe (DINOv2 S/14)	94.4 (12)	65.3 (5)	95.6 (11)	85.6 (12)	85.1 (13)	82.0 (12)
memory, DINOv3 S/16@224	97.0 (5)	60.9 (10)	95.1 (12)	83.1 (14)	96.3 (10)	88.1 (10)
memory, DINOv3 B/16@224, k_{nn}=3	97.0 (6)	61.5 (9)	96.0 (10)	86.1 (9)	96.6 (8)	89.4 (8)
memory, DINOv3 B/16@336	96.7 (8)	56.3 (11)	96.4 (7)	87.5 (8)	96.2 (11)	87.4 (11)
memory, DINOv3 H+/16@224	96.6 (9)	69.3 (2)	96.5 (6)	88.0 (7)	97.4 (7)	92.0 (4)
memory, B/16+H+/16 (reported)	97.1 (4)	64.7 (6)	97.0 (5)	89.4 (5)	97.6 (5)	91.9 (5)
memory, B/16+H+/16, no calibration	97.2 (3)	66.0 (4)	97.0 (4)	89.4 (4)	97.7 (3)	92.2 (3)
memory + background masked out (AnomalyDINO)	4.5 (16)	63.4 (8)	32.1 (16)	89.3 (6)	97.6 (6)	91.9 (7)
memory + query masking	93.3 (14)	67.6 (3)	97.0 (2)	89.5 (2)	97.7 (4)	91.9 (6)
memory + map smoothing (PatchCore)	97.2 (2)	64.6 (7)	97.0 (3)	89.4 (3)	97.8 (2)	92.4 (2)
AnomalyCLIP (Zhou et al., 2024), released checkpoint (zero-shot, trained on MVTec AD)	93.9 (13)	38.3 (15)	95.0 (13)	83.3 (13)	92.6 (12)	74.8 (13)
MVFA (Huang et al., 2024), released checkpoint (annotated 4-shot, one model per dataset)	99.7 (1)	96.3 (1)	97.3 (1)	91.0 (1)	99.0 (1)	95.9 (1)
Kendall τ	0.27		0.75		0.87	
pairs that flip	44		15		8	

A prediction, then a measurement. The identity turns a published number into a prediction about a quantity nobody has reported. From MVFA's published liver CT pixel AUROC of 99.8 at k=8 (Huang et al., 2024) and w=0.94, Table 14 certifies $A_{\mathcal{F}} \geq 96.5$ before any map is looked at. We then ran the authors' released 4-shot checkpoints with their code, supports and zero-/few-shot fusion, and scored the maps on the same pixel sample and partition as every other row (Table 2): A=99.7, whose own bound is 94.7, and $A_{\mathcal{F}}$=96.3, above that bound and within a fraction of a point of the one the published number implied; the bound is tight because $A_{\mathcal{B}}$ is at ceiling. MVFA (annotated supports, one model per dataset) leads under both metrics; its whole-image lead over the best label-free arm is under two points (99.7 against 97.2), its foreground lead 96.3 against 69.3: reporting $A_{\mathcal{F}}$ makes the case for such a method *better*, not worse. The other direction is as easy to find. AnomalyCLIP (Zhou et al., 2024), from its released checkpoint, scores 93.9 on liver CT and $A_{\mathcal{F}}$=38.3: below chance inside the organ, behind a number a reader would take for a working detector.

In Table 2 the maps are the authors' own, and each checkpoint's own evaluation reproduces from them. **Not an artefact of the budget.** At $k \in \{4, 8, 16\}$ (Table 20, Appendix G) liver CT disagrees at every budget (τ=0.07, 0.07, 0.07) and breast ultrasound agrees at every budget ($\tau \geq 0.87$): w governs the disagreement, not k. **And outside medicine.** The same detectors on 27 industrial categories (MVTec AD and VisA, k=16 unlabelled normals, same code path; Table 21, Appendix G) show when the two columns can and cannot come apart. The inflation is there: on the objects the model-free predictor scores 82.9 under A and 63.2 under $A_{\mathcal{F}}$, and the winner changes on 6 of 27 categories. But the frozen-prior memories localise well inside these objects (the best reaches $A_{\mathcal{F}}$=97.7 on average, against 85.1 for the weakest detector), so among them the two rankings largely agree (median τ=0.86 on objects, 1.00 on textures). Eq. 1 says why: two detectors change order only if they differ in $A_{\mathcal{B}} - A_{\mathcal{F}}$, which requires foreground discrimination to be weak relative to background separation. On the medical benchmarks that is every detector; on MVTec it is the weak ones. The re-ranking is a property of how much of the reported number the background carries, not of medicine.

6 A REFERENCE BASELINE REPORTED BOTH WAYS

The audit is not an argument against any method, so we close with a detector reported under both metrics in the most frugal setting we can state: k *unlabelled* normal images per dataset, no anomaly ever seen, no training, one model for nine datasets spanning six modalities (Appendix E). Both

dominant families spend more: vision–language methods adapt CLIP on *annotated* supports with one model per dataset (Huang et al., 2024; Jeong et al., 2023; Chen et al., 2023; Shiri et al., 2025), and reconstruction methods train a decoder per support set (Guo et al., 2025b). Ours is a nearest-neighbour patch memory over frozen DINOv3 tokens in the SPADE/PaDiM/PatchCore line (Cohen & Hoshen, 2020; Defard et al., 2021; Roth et al., 2022; Damm et al., 2025): two frozen taps (ViT-B/16 at 336px, ViT-H+/16 at 224px), cosine distance to the nearest memory entry, the two maps standardised per class and averaged, image score the mean of the top 5% of the map (Appendix M). A trained reconstruction decoder fused with it is discarded by validation (Appendix K).

Table 22 (Appendix H) reports it at $k \in \{4, 8, 16\}$ with two support draws. At k=16 it averages 87.8 image AUROC over nine datasets; on the six mask datasets its whole-image pixel AUROC is 93.8 and its tissue-restricted AUROC 86.5 at k=8, against 63.0 for the blur. Under A its best dataset is liver CT (96.9); under $A_{\mathcal{F}}$ liver CT is its worst (64.1), barely above the mask. On the BMAD six it leads the AnomalyDINO recipe at the same k by 7.2 image-AUROC points and trails methods given k annotated supports and one model per dataset by 13.3 points (Table 23, Appendix I). It is a reference point under a stated budget, not a state of the art; Section 5 re-ranks its maps.

7 Related work

Critiques of pixel AUROC. AUPIMO (Bertoldo et al., 2024) and UniAS (Ma et al., 2025) argue that the metric flatters industrial segmentation and propose stricter per-image and region-level scores; Meissen et al. (2021) show plain intensity thresholding beating trained models on brain MRI, and Meissen et al. (2022) that the residual error is a poor anomaly score; MedIAnomaly (Cai et al., 2025) recommends benchmarking against such baselines, and the BMAD authors warn that a high pixel AUROC there need not reflect an ability to detect anomalies (Bao et al., 2024); a position paper on tabular AD makes the same point in general (Röchner et al., 2025). The metric-pitfall literature (Maier-Hein et al., 2024; Reinke et al., 2024) recommends choosing metrics by problem fingerprint; the composition of the negative class is a fingerprint property pixel AUROC does not see. Restricting a score to a region of interest is an old idea; what the identity adds is the exact relation between the restricted and unrestricted numbers, the bound each places on the other, and the measurement that they order methods differently. The recommendation, report $A_{\mathcal{F}}$ beside A, is compatible with any stricter metric these papers propose and cheaper than all of them.

Few-shot medical anomaly detection. The annotated few-shot protocol is set by MVFA (Huang et al., 2024) on the BMAD suite (Bao et al., 2024) and followed by MediCLIP (Zhang et al., 2024), MadCLIP (Shiri et al., 2025) and Spatial-FAD (Miao et al., 2026), all CLIP-based with one model per dataset. Label-free full-shot baselines on the same suite are MKD (Salehi et al., 2021), PatchCore (Roth et al., 2022) and RD4AD (Deng & Li, 2022). Training-free retrieval over frozen features is AnomalyDINO (Damm et al., 2025) industrially, MuSc (Li et al., 2024) zero-shot, and concurrently HiMatch-AD (Huo et al., 2026) medically with the whole normal pool at inference.

8 Limitations and conclusion

The main limitation is partition dependence: different foreground definitions change $A_{\mathcal{F}}$ while preserving $A = wA_{\mathcal{B}} + (1-w)A_{\mathcal{F}}$. On liver CT, however, this sensitivity is small, with the guarantee changing by only 2.6 points across 6 rules. Re-ranking is also limited to detectors runnable under a common protocol; others can only be bounded (Appendix Q).

Whole-image pixel AUROC is a mixture of two measurements with weights set by the framing of the data, and on the datasets that report the highest numbers the weight on localisation is the smallest. A predictor containing no model comes within three points of the best published method where the frame is emptiest; a published 97% there certifies chance inside the organ; and the detectors the metric ranks first are not the ones that find the lesion best. Whole-image AUROC is safe to read on a benchmark exactly to the extent that $A_{\mathcal{F}}$ is already high for the methods reported on it, a per-benchmark empirical fact, as change detection shows, that the metric cannot certify on its own. Concretely, a paper reporting pixel AUROC should report three more things: w under a stated partition (one pass over the masks), $A_{\mathcal{F}}$ beside A (free once the maps exist), and the model-free floor (one pass over the inputs, no model). Where w is small the extra columns say so; where it is large they say what the headline number was made of.

AI USE STATEMENT

Anthropic's Claude was used solely for language editing and polishing. It was not used to generate research ideas, data, experimental results, or scientific conclusions. The authors take full responsibility for the final manuscript.

ETHICS STATEMENT

This study used only public, previously published, de-identified imaging datasets; no new human data were collected and no ethical approval was required. The audit argues that a widely reported metric overstates localisation on medical benchmarks; we see no harm in that beyond the discomfort of re-reading published numbers, ours included.

REPRODUCIBILITY STATEMENT

Every number in the text is emitted by a generator script from a results file, and the identity is checked numerically on every row it is applied to (Appendix A). The audit needs only the public datasets and a CPU: one file takes images, masks and optional score maps and returns w, $A_{\mathcal{B}}$, $A_{\mathcal{F}}$ and the residual under a stated partition. The re-ranking protocol and every detector's configuration are in Appendix F, the baseline's configuration in Appendix O, its selection procedure in Appendix M, and three evaluation-pipeline defects that would silently corrupt a reproduction in Appendix P. Code, the results files behind every table, and the support-set draws are provided as supplementary material.

REFERENCES

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4042–4053, 2024.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Joao P. C. Bertoldo, Dick Ameln, Ashwin Vaidya, and Samet Akçay. AUPIMO: Redefining visual anomaly detection benchmarks with high speed and low tolerance. In *British Machine Vision Conference (BMVC)*, 2024.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. MedIAnomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, 2025.

Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

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

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pp. 475–489. Springer, 2021.

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.

Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7388–7398, 2022.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 98–107, 2022.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Hongen Liao, and Huiqi Li. One dinomaly2 detect them all: A unified framework for full-spectrum unsupervised anomaly detection. *arXiv preprint arXiv:2510.17611*, 2025b.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11375–11385, 2024.

Jiayu Huo, Jingyuan Hong, Meng Zhou, Liyun Chen, and Le Zhang. HiMatch-AD: DINOv3-driven hierarchical matching for training-free medical anomaly detection. *arXiv preprint arXiv:2606.22556*, 2026.

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

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *International Conference on Learning Representations (ICLR)*, 2024.

Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Wenxin Ma, Qingsong Yao, Xiang Zhang, Zhelong Huang, Zihang Jiang, and S. Kevin Zhou. Towards accurate unified anomaly segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21:195–212, 2024.

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Challenging current semi-supervised anomaly segmentation methods for brain MRI. In *MICCAI Brain Lesion (BrainLes) Workshop*, 2021.

Felix Meissen, Benedikt Wiestler, Georgios Kaissis, and Daniel Rueckert. On the pitfalls of using the residual error as anomaly score. In *Medical Imaging with Deep Learning (MIDL)*, 2022.

Juzheng Miao, Yuchen Yuan, Cheng Chen, and Pheng-Ann Heng. Bridging vision foundation model priors with CLIP for spatial-aware few-shot anomaly detection in medical images. *arXiv preprint arXiv:2609.12454*, 2026.

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

Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädsch, Carole H. Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature Methods*, 21:182–194, 2024.

Philipp Röchner, Simon Klüttermann, Kevin Kammler, Franz Rothlauf, Emmanuel Müller, and Daniel Schlör. We need to rethink benchmarking in anomaly detection. *arXiv preprint arXiv:2507.15584*, 2025.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14902–14912, 2021.

Mahshid Shiri, Cigdem Beyan, and Vittorio Murino. MadCLIP: Few-shot medical anomaly detection with CLIP. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. MediCLIP: Adapting CLIP for few-shot medical image anomaly detection. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Table 3: Every term of the identity on every category audited, on full test splits, under the one background estimator (the region connected to the frame edge within 0.06 of the border's median intensity; the vegetation rule for change detection), for the same model-free predictor (the smoothed distance from the border intensity; the histogram-matched difference map for change detection). w is the fraction of normal pixels the estimator calls background; the whole-image score equals $wA_{\mathcal{B}} + (1-w)A_{\mathcal{F}}$ on every row to within 10^{-13} AUROC points. Inflation is $w(A_{\mathcal{B}} - A_{\mathcal{F}})$. Across all 64 categories Spearman ρ=0.89 (p=9.4×10^{-23}) between w and the inflation; per-source correlations in the block headers. Rows ordered by w within each source.

category	n	w	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl.
Medical, the six of the body (6) — ρ=0.83, p=0.06						
LiverCT	1493	0.939	100.0	14.0	94.8	80.8
RESC	1805	0.781	97.2	42.0	85.1	43.0
BrainMRI	3715	0.718	100.0	86.0	96.1	10.1
HeadCT	110	0.226	99.7	73.0	79.0	6.0
BrainTumor	164	0.167	99.6	79.6	82.9	3.3
BUSI	681	0.154	89.5	39.8	47.5	7.7
Colonoscopy polyps (2D medical, every image anomalous) (4) — ρ=1.00, p=0.08						
CVC-ClinicDB	612	0.347	99.9	65.5	77.4	12.0
CVC-ColonDB	380	0.207	99.9	65.1	72.3	7.2
CVC-300	60	0.167	99.9	70.2	75.2	4.9
Kvasir	1000	0.132	99.5	70.3	74.1	3.9
RoadAnomaly (driving, OOD objects) (1)						
RoadAnomaly	60	0.196	97.3	60.9	68.0	7.1
Building change detection (bi-temporal remote sensing) (4) — ρ= − 0.20, p=0.92						
LEVIR-CD	2048	0.233	56.7	59.1	58.5	-0.6
DSIFN-CD	2488	0.114	64.9	56.7	57.6	0.9
WHU-CD	744	0.110	60.6	54.7	55.4	0.7
S2Looking	16000	0.063	78.9	77.8	77.9	0.1
MVTec AD (15) — ρ=0.77, p=6.9×10^{-4}						
leather	124	0.848	92.1	72.3	89.1	16.7
wood	79	0.720	84.0	50.9	74.7	23.8
toothbrush	42	0.713	100.0	43.4	83.7	40.3
screw	160	0.706	99.9	70.1	91.1	21.0
hazelnut	110	0.687	99.2	59.0	86.6	27.6
pill	167	0.609	100.0	47.6	79.5	31.9
metal_nut	115	0.544	99.9	54.0	79.0	25.0
capsule	132	0.522	100.0	69.6	85.5	15.8
transistor	100	0.395	87.2	42.4	60.1	17.7
bottle	83	0.377	100.0	50.5	69.2	18.6
cable	150	0.236	95.3	59.4	67.9	8.5
zipper	151	0.219	99.9	65.8	73.2	7.5
tile	117	0.180	75.9	50.3	54.9	4.6
carpet	117	0.040	74.3	55.1	55.9	0.8
grid	78	0.031	92.1	63.9	64.8	0.9
VisA (12) — ρ=0.48, p=0.11						
pipe_fryum	600	0.888	100.0	50.4	94.4	44.0
macaroni1	1100	0.755	99.8	74.9	93.7	18.8
macaroni2	1100	0.735	100.0	80.2	94.8	14.5
fryum	600	0.716	100.0	78.8	94.0	15.1
pcb3	1106	0.715	95.8	61.7	86.1	24.4
pcb4	1105	0.704	97.5	62.6	87.2	24.6
cashew	600	0.647	99.7	79.1	92.5	13.3
chewinggum	603	0.549	98.7	70.8	86.1	15.3
pcb2	1101	0.544	99.4	81.8	91.4	9.6
pcb1	1104	0.510	99.8	86.8	93.4	6.7
candle	1100	0.500	100.0	45.7	72.8	27.2
capsules	702	0.177	97.5	54.8	62.4	7.6
DTD-Synthetic (12) — ρ=0.97, p=3.9×10^{-7}						
Stratified_154	100	0.994	81.7	52.6	81.5	29.0
Blotchy_099	100	0.700	90.2	73.8	85.2	11.5
Marbled_078	100	0.617	73.4	50.3	64.5	14.2
Matted_069	99	0.608	70.9	46.8	61.5	14.7
Woven_125	99	0.603	71.1	52.6	63.8	11.2
Fibrous_183	100	0.350	78.9	55.9	63.9	8.0
Woven_104	100	0.135	83.3	50.8	55.1	4.4
Woven_001	100	0.029	89.9	52.6	53.7	1.1
Woven_127	160	0.013	83.2	61.0	61.3	0.3
Perforated_037	100	0.009	92.6	48.6	48.9	0.4
Woven_068	130	0.005	74.4	57.9	58.0	0.1
Mesh_114	116	0.003	78.0	52.4	52.5	0.1
DAGM (10) — ρ=0.95, p=2.3×10^{-5}						
Class1	575	0.734	47.8	20.0	40.4	20.4
Class7	1150	0.345	93.6	52.4	66.6	14.2
Class10	1150	0.343	85.1	45.9	59.4	13.4
Class3	575	0.304	81.5	42.7	54.6	11.8
Class9	1150	0.280	88.7	65.7	72.2	6.4
Class5	575	0.249	77.2	48.8	55.9	7.1
Class8	1150	0.247	83.3	39.8	50.6	10.8
Class2	575	0.041	80.1	63.3	64.0	0.7
Class6	575	0.005	89.4	73.5	73.6	0.1
Class4	575	0.001	65.2	55.5	55.5	0.0

A EVERY TERM OF THE IDENTITY, ON EVERY CATEGORY

Table 3 lists every category of every dataset family audited, on full test splits under the one estimator of Section 2 (the change-detection rows use the vegetation partition of Appendix C): w, $A_{\mathcal{B}}$, $A_{\mathcal{F}}$, the whole-image score, and the inflation; the residual of Eq. 1 is below 10^{-13} AUROC points on every row. Two regimes are visible. Wherever the estimator finds a background region and the predictor separates it, $A_{\mathcal{B}}$ is in the high nineties and the inflation is w times the difference, which is what Fig. 3 plots against the ceiling $100\,w$: scans, endoscopy, objects on a tray, and the uniform textures whose base tone the rule calls background. Where the rule finds no background (w near zero: MVTec's carpet and grid, the densest DTD weaves) there is nothing to inflate, and where the background is not separable to the predictor (change detection, $A_{\mathcal{B}}\approx0.6$) the inflation stays near zero however large w is. Both regimes are the same identity.

A held-out family. Every category above entered the analysis before the paper's claims were written. As a check of the stronger claim of Section 4, we then audited one family that had not: ISIC 2018 dermoscopy (Task 1 validation, 100 images, every image a lesion, public masks), a modality with no black frame but a dark vignette that the border rule finds (w=0.43). The isotonic fit on the 64 categories of Table 3 predicts an inflation of 16.2 points from that w alone (the fit's mean would say 13.3); the measurement is 17.3 (Table 4). The blur's $A_{\mathcal{F}}$ there is 48.7, below chance inside the skin as inside the liver.

Table 4: Held-out replication. A family audited only after every number above was fixed: ISIC 2018 dermoscopy (Task 1 validation, 100 images, every image a lesion, public masks), the same model-free predictor and border rule. The inflation is predicted from w alone by the isotonic fit of Section 4 on the 64 categories of Table 3, before it is measured.

category	n	w	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl. measured	infl. predicted
ISIC2018	100	0.433	88.6	48.7	66.0	17.3	16.2 (isotonic), 11.6 (log-linear)

Table 5: A trained road-scene detector through the identity. The model-free predictor of Table 3 beside the standard baseline of the road-anomaly literature, maximum-softmax probability (and predictive entropy) over a Cityscapes-trained SegFormer-B2 (Xie et al., 2021), released weights, no outlier supervision, on RoadAnomaly (60 frames) / SMIYC RoadAnomaly21 (val) (10 frames) / SMIYC RoadObstacle21 (val) (30 frames). Border-rule partition (the rule finds the sky); void pixels excluded; same w, same pixels per set; the residual of Eq. 1 is below 10^{-10} on every row.

Set	Score	w	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl.
RoadAnomaly	model-free predictor	0.196	97.3	60.9	68.0	7.1
	MSP, SegFormer-B2	0.196	80.6	68.4	70.8	2.4
	entropy, SegFormer-B2	0.196	80.9	68.9	71.2	2.3
SMIYC RoadAnomaly21 (val)	model-free predictor	0.144	97.8	63.2	68.1	5.0
	MSP, SegFormer-B2	0.144	78.3	73.0	73.7	0.8
	entropy, SegFormer-B2	0.144	78.6	73.7	74.4	0.7
SMIYC RoadObstacle21 (val)	model-free predictor	0.105	99.1	77.7	79.9	2.3
	MSP, SegFormer-B2	0.105	99.2	97.6	97.7	0.2
	entropy, SegFormer-B2	0.105	99.2	97.6	97.8	0.2

Where the segmenter's $A_{\mathcal{B}}$ comes from, on RoadAnomaly: the median MSP is 4×10^{-5} on sky pixels and 0.046 on anomaly pixels, and 7.6% of anomaly pixels score below the sky median. The segmenter is confident on the sky; $A_{\mathcal{B}}$=80.6 is set by the anomaly pixels it is also confident about, the failure mode MSP is known for, not by the sky scoring high.

The audit is released as a single file with no dependency beyond NumPy, SciPy and Pillow. It takes a directory of images, a directory of masks and optionally a directory of score maps, and returns every term of Eq. 1 under a stated partition. Run on one MVTec AD category with no score maps, so that it scores its own model-free predictor, it prints:

```
$ python auroc_decompose.py --images hazelnut/test/hole \
                            --masks hazelnut/gt/hole
images 18  w 0.6706  A 0.8361  A_B 0.9903  A_F 0.5223
inflation 0.3138  residual 0.0e+00  guarantee A_F >= 0.5025
```

Two thirds of that category's normal pixels are background, the predictor separates them almost perfectly, and what is left inside the object is chance. Pointing the same command at a directory of a method's score maps returns the same decomposition for that method.

B THE BACKGROUND CONTROL, PER DATASET

Fig. 6 draws the blur rows of Table 1 per dataset. Table 6 sweeps the partition: the border tolerance over 0.03–0.10 and an intensity threshold over 0.05–0.20, all on the full splits, accumulated in one pass. Table 7 repeats Table 1 in full under the threshold at 0.1, the rule the baseline's tissue-restricted metrics use. The identity holds to the same precision under every partition; the terms move. On liver CT the threshold keeps slightly more of the frame and the blur's $A_{\mathcal{F}}$ reads 12.7; on retinal OCT the threshold counts the speckled vitreous as tissue (72.8% foreground against the border rule's retina-only $\mathcal{F}$), which lowers w from 0.78 to 0.28 and moves the inflation accordingly; on breast ultrasound the threshold assigns 25% of the lesion pixels, the dark hypoechoic ones, to the background, and since $A_{\mathcal{F}}$ scores all of them, that dataset's $A_{\mathcal{F}}$ falls to 47.5 and its inflation rises to -0.0, out of the order the other five keep (ρ=0.77, p=0.10 over the six). Under that partition the field's convention

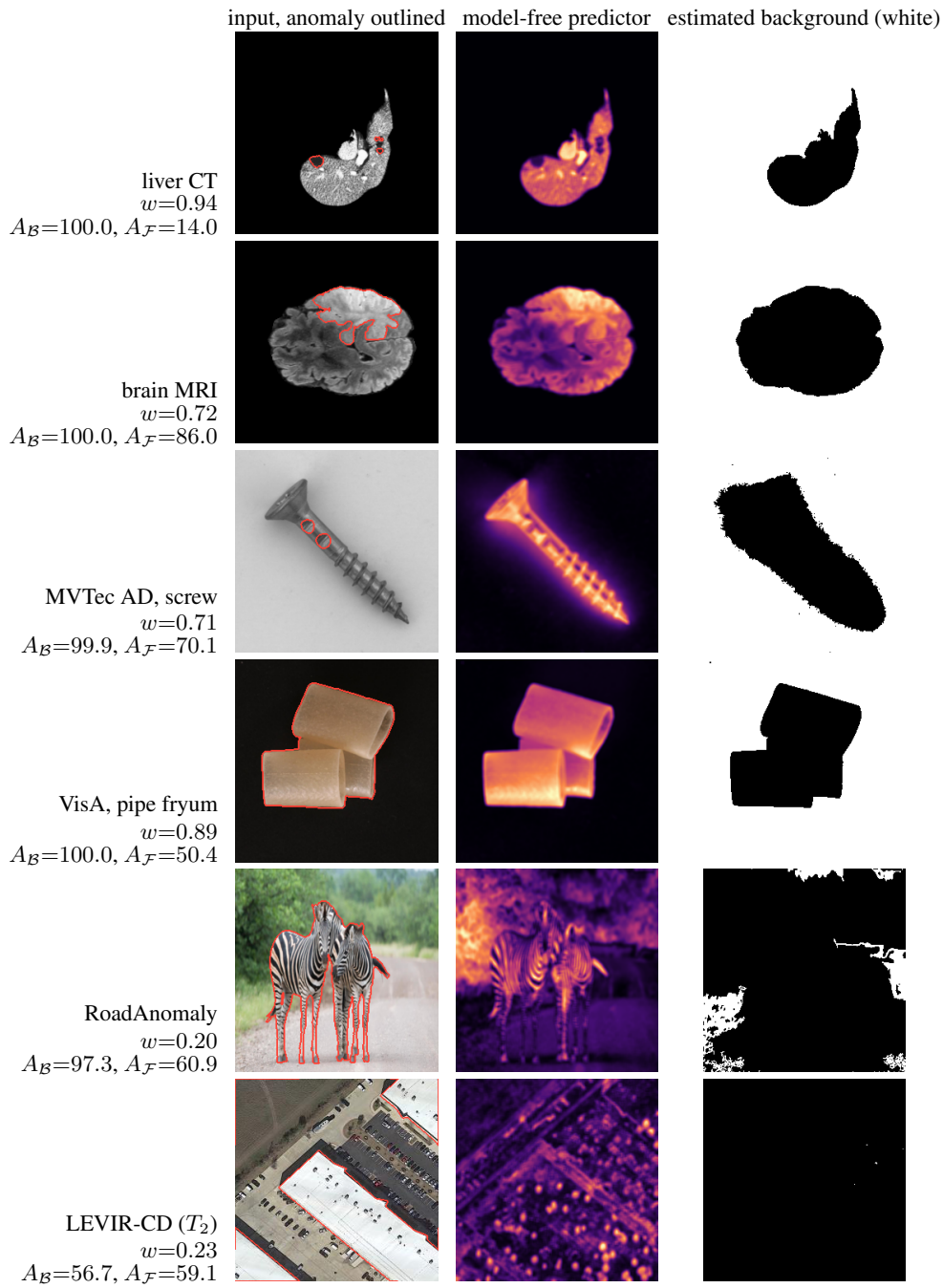


Figure 5: The same audit on six families, one image each. Columns: the input with the anomaly outlined; the model-free predictor (smoothed distance from the border intensity; for change detection the smoothed difference of the histogram-matched pair); the estimated background (border rule; for change detection, vegetation in both epochs). The numbers at the left are the category's terms from Table 3. Where the background is black, a tray or sky, the predictor separates it ($A_{\mathcal{B}}$ near 100) and the whole-image number is mostly w; on the change-detection tile the vegetation is not separable to a difference map ($A_{\mathcal{B}}$=56.7) and there is nothing to inflate.

of restricting both classes to the foreground also drops those lesion pixels, and differs from $A_{\mathcal{F}}$ by 4.1 points there against at most 0.8 elsewhere.

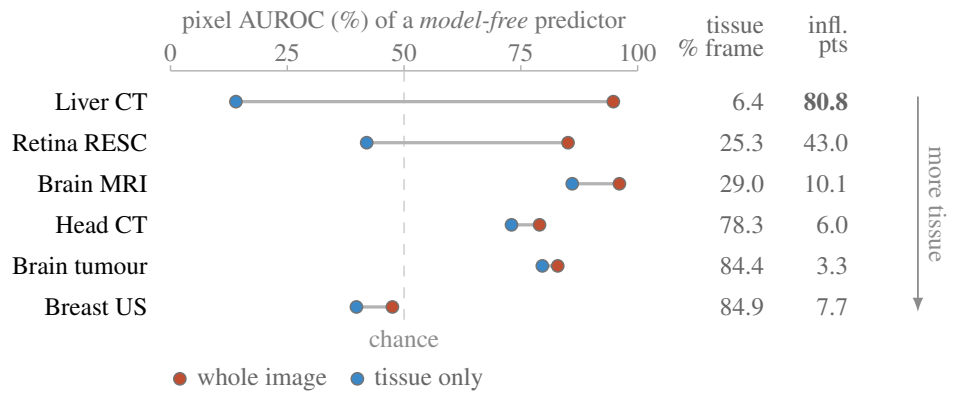


Figure 6: The blur rows of Table 1, drawn per dataset. Each row is the Gaussian blur of the input scored over the whole frame (red) and against foreground normals only (blue); the bar is the inflation. Rows are sorted by foreground fraction, and the inflation follows that order.

Table 6: Sensitivity of the audit to the foreground partition, on the full test splits of the six medical datasets. Each row is one partition rule; the cells give the blur's $A_{\mathcal{F}}$ per dataset, then the rank correlation between w and the inflation over the six, the largest fraction of anomalous pixels any dataset's rule puts in the background, and liver CT's w. The identity holds under every row; what moves is which pixels count as foreground, and only breast ultrasound (dark lesions) and retinal OCT (dark vitreous) move by more than a few points.

Partition	Liver CT	Retina RESC	Brain MRI	Head CT	Brain tumour	Breast US	ρ (p)	max leak	w_{liver}
border, tol 0.03	15.1	64.2	86.1	73.8	80.3	45.9	0.94 (0.02)	1.1%	0.939
border, tol 0.06 (used)	14.0	42.0	86.0	73.0	79.6	39.8	0.83 (0.06)	11.8%	0.939
border, tol 0.10	12.9	28.6	85.8	72.7	79.0	31.6	0.94 (0.02)	34.2%	0.940
intensity $>$ 0.05	14.1	85.7	85.9	75.7	80.2	47.5	0.94 (0.02)	14.4%	0.939
intensity $>$ 0.10	12.7	81.3	85.5	74.1	79.0	47.5	0.77 (0.10)	24.5%	0.940
intensity $>$ 0.20	10.9	24.8	83.4	71.5	76.8	44.9	0.77 (0.10)	46.6%	0.942

The estimator against a ground-truth organ mask. Liver CT is the dataset the argument leans on, and its test slices are cut from ten LiTS volumes (Bilic et al., 2023) whose liver labels are public. We re-identified every slice: the lesion mask against the LiTS tumour label (IoU ≥ 0.95), then the image against the LiTS slice by Spearman correlation inside the liver label (≥ 0.9); the volume index and slice index in BMAD's file names turn out to be LiTS's own, with no transform. 644 of 660 anomalous test slices pass (the rest fail on lesion-mask IoU, i.e. their masks differ from the LiTS label); the normal test slices align to the same ten volumes (808 of 833), so the ground-truth partition covers the liver test set. On the passing slices the border-rule foreground agrees with LiTS's liver segmentation at Dice 0.98: the benchmark's slices are masked to the organ, and the rule recovers that mask. Replacing the estimator by the ground-truth mask moves w from 0.939 to 0.940 and the blur's $A_{\mathcal{F}}$ from 14.0 to 11.9 (the mask predictor's from 49.9 to 50.1); at the ground-truth w a published 97% certifies $A_{\mathcal{F}} \geq 49.8$. The 12 re-ranking arms of Table 18 re-evaluated on those slices under both partitions (Table 11) keep their $A_{\mathcal{F}}$ order (τ=1.00 between the two $A_{\mathcal{F}}$ rankings), and A against $A_{\mathcal{F}}$ under ground truth gives τ=0.12 against 0.12 under the estimator on the same images.

AUPRO inherits the identity on its x axis. Per-region overlap is a function of the anomalous pixels only; the false-positive rate on the axis it is integrated over is a mean over the normal pixels, so it splits exactly as $\mathrm{FPR}(t) = w\,\mathrm{FPR}_{\mathcal{B}}(t) + (1-w)\,\mathrm{FPR}_{\mathcal{F}}(t)$. When the background is rejected, the reported curve is the foreground curve compressed by $(1-w)$ along x: the standard AUPRO up to FPR 0.3 reads the foreground curve out to $0.3/(1-w)$, and where $1-w < 0.3$ every foreground normal is already a false positive at whole-image FPR $1-w$, after which the integrand sits at the overlap reached with the foreground exhausted. That prediction, computed from the foreground curve alone, matches the blur's reported number to a median of 0.0 points over the six datasets and at worst 2.9 points (Retina RESC, where $A_{\mathcal{B}}$ is 97 rather than 100 and the background is not fully rejected); the mask predictor's rows behave the same way (Table 12). The consequence is larger than for AUROC: on liver CT the blur's AUPRO is 83.2 as reported and 2.1 against foreground normals, because at w=0.94 four fifths of the reported integral lies where every tissue pixel is already flagged.

Table 7: Every term of the identity on the six medical mask datasets, for two predictors that contain no model, with the foreground estimated by thresholding the input at 0.1, the rule the model's tissue-restricted metric uses, on the full test splits. w is the fraction of normal pixels that are background; $A_{\mathcal{B}}$ and $A_{\mathcal{F}}$ are the AUROC of anomalous pixels against background and against foreground normals; the whole-image column is what the field reports and equals $wA_{\mathcal{B}} + (1-w)A_{\mathcal{F}}$ on every row to within 10^{-10}. The inflation is the whole-image score minus $A_{\mathcal{F}}$, i.e. $w(A_{\mathcal{B}} - A_{\mathcal{F}})$; the last column is the exchange rate $1/(1-w)$. Spearman ρ=0.77 (p=0.10) between w and the blur's inflation. Rows are ordered by w.

	fg.		Gaussian blur of the input				binary foreground mask				
Dataset	% frame	w	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl.	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl.	$\frac{1}{1-w}$
Liver CT	6.3	0.940	100.0	12.7	94.8	82.1	99.8	49.8	96.8	47.0	16.8
Brain MRI	28.0	0.728	100.0	85.5	96.1	10.5	100.0	50.0	86.4	36.4	3.7
Retina RESC	72.8	0.285	94.6	81.3	85.1	3.8	98.7	48.7	62.9	14.2	1.4
Brain tumour	81.7	0.195	99.0	79.0	82.9	3.9	99.7	49.7	59.5	9.7	1.2
Head CT	81.3	0.194	99.6	74.1	79.0	4.9	99.7	49.7	59.4	9.7	1.2
Breast US	82.0	0.174	47.5	47.5	47.5	-0.0	87.7	37.7	46.4	8.7	1.2

Table 8: Which metrics the background reaches, and which does not. The same Gaussian blur scored over the whole frame and against foreground normals only (all anomalous pixels kept), on the full test splits, for four pixel metrics. Pixel AUROC moves by the identity's $w(A_{\mathcal{B}} - A_{\mathcal{F}})$. AP and F1-max barely move: a trivially-separable background sits at the bottom of the score, below every real operating point, so removing it changes almost no precision–recall pair. The per-image pixel AUROC (mean over anomalous images) obeys the identity image by image with that image's own w, and is inflated by the same mechanism as the whole-set score.

	pixel AUROC		pixel AP		F1-max		per-image AUROC	
Dataset	whole	fg	whole	fg	whole	fg	whole	fg
Liver CT	94.8	14.0	3.4	3.4	11.2	11.4	94.5	15.2
Retina RESC	85.1	42.0	15.2	16.0	27.9	32.7	88.5	48.3
Brain MRI	96.1	86.0	19.8	19.8	30.1	30.1	97.6	91.5
Head CT	79.0	73.0	9.5	9.5	22.0	22.0	80.5	74.9
Brain tumour	82.9	79.6	26.6	26.6	36.9	36.9	85.9	82.9
Breast US	47.5	39.8	8.0	8.1	16.3	18.7	47.3	39.8

Table 8 gives four pixel metrics for the blur under the border partition, whole-frame and foreground. Table 9 gives the six datasets under the field's convention (both classes inside the threshold tissue) beside the baseline at k=8, so the baseline can be read the same two ways; its model-free columns are the ones a reader of the tissue-restricted columns of Table 22 should compare against. Lesions occupy 4.1% (brain), 7.1% (liver), 5.9% (retina) and 6.0% (breast) of tissue, so the tissue-restricted AP of a random scorer is that same figure. Table 10 is the imbalance control of Section 3, on the 300-image subsample per class.

C CHANGE DETECTION THROUGH THE IDENTITY

Bi-temporal building change detection is scored on the same kind of object, per-pixel masks over an image, but by a community that reports F1 and IoU. We audit it as a dense-evaluation neighbour. Each test tile pair (LEVIR-CD, WHU-CD and DSIFN-CD at their native 256×256; S2Looking's 1024×1024 tiles cut into sixteen) is scored by the classic model-free change map, the smoothed absolute grey-level difference after histogram-matching the second epoch to the first. Nothing in a tile is connected to a frame edge at a constant intensity, so the border rule does not apply; the model-free partition here calls a pixel background if it is vegetation in both epochs by the excess-green index $2G-R-B > \tau$, on the grounds that building change cannot occur where only vegetation was ever present. Table 13 gives every term at three τ. The background is not free here: seasonal and illumination differences between epochs are as large on vegetation as on buildings to a pixel-difference map, so $A_{\mathcal{B}}$ is far from 1 and the inflation is small even where w is large. The identity

Table 9: The reference baseline read the same two ways as the model-free predictors, on the six mask datasets, ordered by the fraction of the frame that is tissue. Both blocks use the field's tissue-restricted convention (both classes inside the thresholded tissue), which is why the model-free columns differ slightly from $A_{\mathcal{F}}$ in Table 1; the predictors and the splits are the same. The baseline is at k=8. Its whole-image column ranks liver CT first and its tissue column ranks it last, which is the re-ranking of Section 5 seen on one detector.

			no model		SUMMA (ours)	
Dataset	tissue % frame	lesion % tissue	intensity whole	tissue only	whole image	tissue only
Liver CT	6.3	7.1	94.8	12.8	96.9	64.1
Brain MRI	28.0	4.1	96.1	85.5	96.7	88.7
Retina RESC	72.8	6.7	85.1	82.1	97.2	96.2
Head CT	81.3	4.9	79.0	74.5	96.3	95.8
Brain tumour	81.7	7.9	82.9	79.5	93.4	92.5
Breast US	82.0	7.9	47.5	43.4	80.7	81.5
Mean			**80.9**	**63.0**	**93.5**	**86.5**

Table 10: Class imbalance is not what inflates the metric. The same model-free predictor (Gaussian blur) scored three ways: as reported; with normal pixels subsampled at random to equal the anomalous pixels (mean $\pm$ std over five draws), which removes the imbalance and preserves w; and with only the foreground normals kept, which preserves most of the imbalance and sets w to zero. AUROC is a probability over (anomalous, normal) pairs, so the first operation cannot move it beyond sampling noise, while the second moves it by the full inflation.

Dataset	normal : anomalous	as reported	classes balanced	background removed
Liver CT	219 : 1	94.8	94.8 ± 0.0	14.4
Retina RESC	19 : 1	85.1	85.0 ± 0.0	42.1
Brain MRI	86 : 1	96.1	96.0 ± 0.0	85.6
Head CT	24 : 1	79.0	79.0 ± 0.1	73.0
Brain tumour	14 : 1	82.9	82.9 ± 0.0	79.6
Breast US	11 : 1	47.5	46.6 ± 0.0	38.6

is exact on every row; the lesson is that whether a benchmark's whole-image AUROC is mostly w depends on whether its background is trivially separable, which is a property of the framing and the predictor together, and which the audit makes measurable.

A trained detector, not only the model-free predictor. The paragraph above establishes that a pixel-difference map does not find the vegetation background trivially separable. A reviewer's fair objection is that a *trained* building-change detector might: it has seen thousands of examples of "this is vegetation, not a building" and could reject it with near-certainty, which would put $A_{\mathcal{B}}$ near 1 even where the model-free predictor's is not, restoring room for inflation. We tested this rather than assert it either way. Running BIT (Chen et al., 2021), the official released LEVIR-CD checkpoint (best validation accuracy 0.947, matching the authors' reported figure), through the authors' own code over all 2048 LEVIR-CD test tiles, and decomposing its change-probability map on the same excess-green partition (τ=0.05, w=0.233): $A_{\mathcal{B}}$=99.5, $A_{\mathcal{F}}$=99.4, whole-image $= 99.4$, inflation 0.0 points, identity residual $0.0e + 00$. The objection is half right and half wrong. $A_{\mathcal{B}}$ *is* near ceiling for a trained detector, as predicted; but so is $A_{\mathcal{F}}$, because BIT is simply a strong detector everywhere on this benchmark, not only against vegetation. The identity's prediction is confirmed in the direction that matters: inflation requires $A_{\mathcal{B}}$ to exceed $A_{\mathcal{F}}$, and for a detector whose foreground discrimination is itself excellent there is no gap left to inflate. This is the cleanest contrast with medical AD in the paper. On liver CT, Table 18 shows every detector we ran, trained or training-free, scoring $A_{\mathcal{F}}$ under 70; the reported whole-image number is high regardless, because $A_{\mathcal{B}} \approx 1$ there too, but for a different reason — the organ mask is easy, not the lesion. On LEVIR-CD, a real detector closes that gap by being good at the actual task. Whole-image AUROC is safe to read on a benchmark exactly to the extent that $A_{\mathcal{F}}$ is already high for the methods reported on it, which is a per-benchmark empirical fact and not something the metric can tell you on its own. We do

Table 11: The re-ranking arms on liver CT under the estimated partition and under LiTS's ground-truth liver mask, on the 1452 re-identified test slices and the same pixel sample. τ between the two $A_{\mathcal{F}}$ rankings is 1.00; between A and $A_{\mathcal{F}}$ it is 0.12 under the estimator and 0.12 under ground truth.

	estimator (w=0.939)		ground truth (w=0.940)	
Detector	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$
CLIP zero-shot, prompts only (VAND/APRIL-GAN branch)	73.8	53.1	73.8	53.0
PatchCore features (WRN50-2), k supports	95.2	50.6	95.2	50.5
AnomalyDINO recipe (DINOv2 S/14)	94.7	65.4	94.7	65.4
memory, DINOv3 S/16@224	97.0	61.0	97.0	60.8
memory, DINOv3 B/16@224, k_{nn}=3	96.9	61.4	96.9	61.2
memory, DINOv3 B/16@336	96.7	56.4	96.7	56.3
memory, DINOv3 H+/16@224	96.6	69.2	96.6	69.2
memory, B/16+H+/16 (reported)	97.1	64.7	97.1	64.7
memory, B/16+H+/16, no calibration	97.1	65.9	97.1	65.9
memory + background masked out (AnomalyDINO)	4.5	63.3	4.5	64.0
memory + query masking	93.1	67.5	93.1	67.2
memory + map smoothing (PatchCore)	97.1	64.6	97.1	64.5

Table 12: AUPRO through the identity. Per-region overlap is a function of the anomalous pixels only; the false-positive rate on its x axis splits exactly as $\mathrm{FPR}(t) = w\,\mathrm{FPR}_{\mathcal{B}}(t) + (1-w)\,\mathrm{FPR}_{\mathcal{F}}(t)$. For the two model-free predictors on the full test splits: the standard AUPRO up to FPR 0.3 over all normal pixels (what is reported), the same with only foreground normals as negatives, and the reported number as the identity predicts it from the foreground curve alone when the background is rejected: the foreground AUPRO read out to $0.3/(1-w)$ if that is below 1, otherwise a $(1-w)$-weighted foreground AUPRO plus a plateau at $\mathrm{PRO}_{\mathcal{F}}(1)$ over the remaining $0.3-(1-w)$ of the axis. Values in %.

		Gaussian blur of the input			binary foreground mask		
Dataset	w	reported	fg only	predicted	reported	fg only	predicted
Liver CT	0.939	83.2	2.1	83.2	89.8	0.0	89.9
Retina RESC	0.781	58.1	10.8	60.9	57.4	0.0	63.4
Brain MRI	0.718	82.0	50.9	82.0	53.0	0.0	53.0
Head CT	0.226	38.1	28.3	38.1	0.0	0.0	0.0
Brain tumour	0.167	51.5	46.9	51.5	0.0	0.0	0.0
Breast US	0.154	10.0	8.0	10.0	0.0	0.0	0.0

not have an independent building-footprint mask to partition by instead of the vegetation heuristic (WHU-CD's own masks are the change labels, not object footprints); this is a limitation of what the audit can test with public data.

D Reading a published number against its floor

A second convention, complementary to the guarantee of Table 14: the floor is the better of the two model-free predictors on a dataset, the headroom is 100 minus the floor, and the fraction of the headroom a method takes is $(A - \text{floor})/(100 - \text{floor})$. Table 16 applies it to published pixel AUROCs. The headroom differs by a factor of four across the three datasets every published few-shot paper reports, so the same raw number means different things, and the full-pool method's ranking of its own datasets inverts between raw score and headroom taken. The floor is computable from the input images alone, and any paper can report its own.

E Benchmark composition

All nine datasets are handled by a *single* baseline model: one set of weights, one forward pass per test image. Support sets are drawn from the training pool of normal images only; the pool size is listed for context but the model never sees more than k images per class.

Table 13: Building change detection through the identity, on the test splits. The predictor is the smoothed absolute grey-level difference of the histogram-matched pair, the classic model-free change map; the partition calls a pixel background if it is vegetation in *both* epochs by the excess-green index $2G-R-B > \tau$, where building change cannot occur. Three τ are shown. $A_{\mathcal{B}}$ is far from 1: seasonal vegetation differences are as large as building differences to a pixel-difference map, so the background is not free here, and the inflation is small even where w is large. That is the identity's prediction, not a failure of it.

Dataset	τ	n	w	$A_{\mathcal{B}}$	$A_{\mathcal{F}}$	whole	infl.
DSIFN-CD	0.02	2488	0.211	61.6	56.5	57.6	1.1
	0.05	2488	0.114	64.9	56.7	57.6	0.9
	0.1	2488	0.041	68.0	57.2	57.6	0.4
LEVIR-CD	0.02	2048	0.488	56.6	60.3	58.5	-1.8
	0.05	2048	0.233	56.7	59.1	58.5	-0.6
	0.1	2048	0.027	59.5	58.5	58.5	0.0
S2Looking	0.02	16000	0.257	77.9	77.8	77.9	0.0
	0.05	16000	0.063	78.9	77.8	77.9	0.1
	0.1	16000	0.002	63.8	77.9	77.9	-0.0
WHU-CD	0.02	744	0.226	59.0	54.3	55.4	1.1
	0.05	744	0.110	60.6	54.7	55.4	0.7
	0.1	744	0.025	64.9	55.1	55.4	0.2

Table 14: What a whole-image pixel AUROC guarantees inside the foreground. Because $A_{\mathcal{B}} \leq 1$, a reported score A implies $A_{\mathcal{F}} \geq (A-w)/(1-w)$; the table evaluates that bound with each dataset's w measured on its full test split. Upper block: a grid of reported values. Lower block: published numbers, with the bound they certify. The bound is the only thing the reported number certifies; it is not an estimate of the foreground score, which can sit anywhere above it.

	Liver CT	Brain MRI	Retina RESC
w (full split)	0.939	0.718	0.781
reported whole-image AUROC → guaranteed $A_{\mathcal{F}}$			
95.0	17.4	82.3	77.2
97.0	50.4	89.4	86.3
98.0	67.0	92.9	90.9
99.0	83.5	96.5	95.4
99.5	91.7	98.2	97.7
99.8	96.7	99.3	99.1
99.9	98.3	99.6	99.5
published			
MVFA (Huang et al., 2024), k=8 annotated	99.8 → 96.5	97.2 → 90.1	99.0 → 95.4
HiMatch-AD (Huo et al., 2026), full pool	98.9 → 82.2	98.9 → 96.0	97.0 → 86.4

Three details affect comparability. First, the chest X-ray test split is subsampled to 3000 images (all 781 normals plus a fixed random draw of 2219 anomalies, seed 133); the full split carries 17,194 images, over half the benchmark, and evaluating it in full would have dominated wall-clock without changing any conclusion. Second, chest X-ray and histopathology ship *no* lesion masks; the data loader supplies an all-ones placeholder mask for anomalies there, so a "pixel AUROC" can still be computed, but it measures the placeholder, not the model, and we exclude all three detection-only datasets from every pixel-level number. Three of the nine ship no validation split, so the configuration was selected on the six that do and applied to them unchanged. Third, head CT and the second brain MRI source ship only 10 normal test images each; an image-level AUROC estimated against ten negatives has a standard error of roughly 0.05, so those two columns are indicative and are not used in any claim that rests on a small difference.

Table 15: The points of Fig. 2: every published pixel AUROC on the three BMAD segmentation datasets at 16 annotated shots, as tabulated in MadCLIP's Table 1 (Shiri et al., 2025) (*M*), and HiMatch-AD's own Table 2 (*H*, full normal pool, training-free), with the $A_{\mathcal{F}}$ each certifies under the dataset's full-split w (Table 14). Published values are transcribed, not re-run; MVFA's k=8 liver CT entry, 99.8, is the number the text quotes and its k=16 entry is 99.7. Bounds below zero are shown as 0.

		Liver CT		Brain MRI		Retina	
Method	src	A	$\geq A_{\mathcal{F}}$	A	$\geq A_{\mathcal{F}}$	A	$\geq A_{\mathcal{F}}$
CFlow-AD (Gudovskiy et al., 2022)	M	92.8	0	93.5	77	93.8	72
RD4AD (Deng & Li, 2022)	M	95.9	32	96.5	88	96.2	83
PatchCore (Roth et al., 2022)	M	96.6	44	97.0	89	96.4	84
MKD (Salehi et al., 2021)	M	96.1	36	89.5	63	86.6	39
DRA (Ding et al., 2022)	M	93.0	0	80.5	31	84.0	27
BGAD (Yao et al., 2023)	M	99.2	88	95.3	83	97.1	87
APRIL-GAN (Chen et al., 2023)	M	99.6	94	96.2	86	98.5	93
MediCLIP (Zhang et al., 2024)	M	99.0	83	98.1	93	94.1	73
MVFA (Huang et al., 2024)	M	99.7	96	97.7	92	99.1	96
MadCLIP (Shiri et al., 2025)	M	99.7	96	98.0	93	99.5	97
HiMatch-AD (Huo et al., 2026)	H	98.9	82	98.9	96	97.0	86

Table 16: Published *pixel* AUROC read against the model-free floor of its dataset. The floor is the better of the two predictors that contain no model, so the headroom is what a method can still win; it differs by a factor of four across these three datasets. AF is the fraction of that headroom a method actually takes. The reading changes: MVFA's brain MRI number looks comparable to its others in raw terms and takes barely a quarter of the headroom, and the full-pool method's ranking of its own three datasets inverts. Our floors come from a subsample of each test split while the published numbers use the full split, so these are indicative — the point is that the floor is computable from the input images alone, and every paper can report its own.

	Liver CT		Brain MRI		Retina RESC	
Method	raw	AF	raw	AF	raw	AF
model-free floor	96.8		95.9		86.1	
headroom	3.2		4.1		13.9	
MVFA (Huang et al., 2024), k=8 annotated	99.8	**93%**	97.2	**31%**	99.0	**93%**
HiMatch-AD (Huo et al., 2026), full pool	98.9	**66%**	98.9	**72%**	97.0	**78%**

F THE RE-RANKING PROTOCOL

Sampling interval for Kendall's τ: re-evaluating the 11 memory and feature arms on six pixel samples per support draw (12 replicates, the other arms at their table values), τ on liver CT stays within 0.12–0.19 (pairs flipping 37–40 of 91) and on breast ultrasound within 0.98–0.98; the two ranges do not overlap (per support draw alone: liver CT 0.05 and 0.09, breast ultrasound 0.96 on both). Ranking the same detectors by pixel AP restricted to foreground normals agrees with the $A_{\mathcal{F}}$ ranking at τ=0.79 (liver CT), 0.79 (retina), 0.87 (brain MRI), 0.82 (head CT), 0.79 (brain tumour) and 0.54 (breast ultrasound), the per-dataset values behind the range quoted in Section 3.

Every detector in Table 18 runs through the same evaluation code (`eval_ensemble.py` in the supplementary material) on the six mask datasets, with k=16 unlabelled normal supports (seeds 0 and 1, averaged) where a support set is used. Maps are evaluated at 256×256; whole-image pixel AUROC, the tissue-restricted variants and every term of the identity are computed on one fixed random sample of 10^7 pixels per class (standard error $\approx 2 \times 10^{-4}$), so that $A = wA_{\mathcal{B}} + (1-w)A_{\mathcal{F}}$ holds on the sample exactly. Detector configurations: DINOv3 memories read four blocks ($\{2,4,6,8\}$ for S/16 and B/16, $\{5,11,16,21\}$ for H+/16) at the stated resolution, k_{nn}=1 unless stated, image score the mean of the top 5% of the map; the AnomalyDINO recipe uses DINOv2 ViT-S/14 at 224px, blocks $\{9,10,11,12\}$, k_{nn}=1 and the top-1% image score of its paper; PatchCore features are WideResNet-50-2 layer2+layer3 with 3×3 neighbourhood averaging, layer3

Table 17: Benchmark composition. One unified model covers all nine. 'Masks' counts anomalous test images with a real lesion mask; the three detection-only datasets are excluded from every pixel-level result.

Dataset	Modality	Normal	Anomalous	Masks
BraTS2021	MRI	640	3075	3075
BTCV+LiTS	CT	833	660	660
RESC	OCT	1041	764	764
OCT2017	OCT	242	726	–
RSNA	X-ray	781	2219	–
Camelyon16	Histopath.	1003	994	–
BUSI	Ultrasound	34	647	647
Head CT	CT	10	100	100
Brain MRI (Kaggle)	MRI	10	154	154
Total		**4594**	**9339**	**5400**

Table 18: Re-ranking under the two metrics. Every row is a detector run through one evaluation code path on the six mask datasets (k=16 normal supports where a support set is used, two draws averaged), with whole-image pixel AUROC A and the foreground term $A_{\mathcal{F}}$ measured on the same pixel sample under the border partition. Small numbers are the rank within the column (1 = best). The last rows give Kendall's τ between the two rankings and the number of detector pairs whose order flips, out of 91. The two model-free predictors are ranked with the detectors: under A the blur outranks real detectors where the frame is empty; under $A_{\mathcal{F}}$ it does not.

	Liver CT		Retina		Brain MRI		Head CT		Brain tum.		Breast US	
Detector	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$
Gaussian blur of the input (no model)	94.8 (10)	14.0 (14)	85.1 (12)	42.0 (14)	96.1 (7)	86.0 (10)	79.0 (12)	73.0 (12)	82.9 (12)	79.6 (12)	47.5 (14)	39.8 (14)
foreground mask (no model)	96.9 (6)	49.9 (13)	83.1 (13)	44.1 (13)	85.9 (12)	50.0 (14)	61.3 (13)	50.0 (14)	58.3 (14)	50.0 (14)	51.9 (13)	44.2 (13)
CLIP zero-shot, prompts only (VAND/APRIL-GAN branch)	74.1 (13)	53.2 (11)	61.8 (14)	71.2 (12)	74.5 (13)	53.7 (13)	55.1 (14)	51.4 (13)	68.2 (13)	66.6 (13)	70.7 (11)	69.8 (11)
PatchCore features (WRN50-2), k supports	95.3 (9)	50.7 (12)	96.4 (8)	88.3 (8)	96.0 (8)	86.1 (9)	92.2 (11)	92.6 (9)	89.0 (10)	88.4 (10)	77.5 (6)	77.6 (6)
AnomalyDINO recipe (DINOv2 S/14)	94.4 (11)	65.3 (4)	85.1 (11)	82.0 (11)	95.6 (10)	85.6 (11)	94.2 (6)	94.2 (7)	91.2 (7)	91.1 (8)	75.0 (9)	75.2 (8)
memory, DINOv3 S/16@224	97.0 (4)	60.9 (9)	96.3 (9)	88.1 (9)	95.1 (11)	83.1 (12)	93.9 (7)	93.4 (8)	87.8 (11)	86.5 (11)	75.8 (7)	75.7 (7)
memory, DINOv3 B/16@224, k_{nn}=3	97.0 (5)	61.5 (8)	96.6 (7)	89.4 (7)	96.0 (9)	86.1 (8)	92.9 (9)	92.2 (11)	91.0 (8)	90.2 (9)	75.1 (8)	74.8 (9)
memory, DINOv3 B/16@336	96.7 (7)	56.3 (10)	96.2 (10)	87.4 (10)	96.4 (6)	87.5 (7)	93.1 (8)	92.2 (10)	92.2 (6)	91.4 (7)	73.9 (10)	73.5 (10)
memory, DINOv3 H+/16@224	96.6 (8)	69.3 (1)	97.4 (6)	92.0 (3)	96.5 (5)	88.0 (6)	96.0 (5)	95.2 (5)	92.6 (5)	91.8 (6)	82.2 (2)	81.8 (2)
memory, B/16+H+/16 (reported)	97.1 (3)	64.7 (5)	97.6 (4)	91.9 (4)	97.0 (4)	89.4 (4)	96.2 (4)	95.5 (3)	94.1 (4)	93.4 (2)	81.0 (5)	80.6 (5)
memory, B/16+H+/16, no calibration	97.2 (2)	66.0 (3)	97.7 (2)	92.2 (2)	97.0 (3)	89.4 (3)	96.3 (2)	95.7 (2)	94.1 (3)	93.4 (3)	81.8 (3)	81.3 (3)
memory + background masked out (AnomalyDINO)	4.5 (14)	63.4 (7)	97.6 (5)	91.9 (6)	32.1 (14)	89.3 (5)	92.5 (10)	95.1 (6)	89.2 (9)	92.8 (5)	82.9 (1)	82.6 (1)
memory + query masking	93.3 (12)	67.6 (2)	97.7 (3)	91.9 (5)	97.0 (1)	89.5 (1)	96.3 (3)	95.4 (4)	94.2 (2)	93.3 (4)	67.2 (12)	66.9 (12)
memory + map smoothing (PatchCore)	97.2 (1)	64.6 (6)	97.8 (1)	92.4 (1)	97.0 (2)	89.4 (2)	96.3 (1)	95.7 (1)	94.2 (1)	93.6 (1)	81.4 (4)	81.0 (4)
Kendall τ (A vs $A_{\mathcal{F}}$)	0.12		0.85		0.71		0.82		0.85		0.98	
pairs that flip (of 91)	40		7		13		8		7		1	

upsampled to layer2's grid and concatenated, keyed by cosine distance in the same memory, image score the maximum of a 16×16 average-pooled map. Background masking removes support tokens whose patch is more than half background under the threshold rule; query masking holds background query tokens at the image's own foreground minimum; smoothing is a Gaussian with σ=4 on the upsampled map. The model-free rows are the blur and mask of Section 3 on the full splits.

G THE RE-RANKING AT OTHER BUDGETS, AND OUTSIDE MEDICINE

MVFA's released checkpoints, against every training-free arm on the three BMAD datasets they cover, are Table 2 in the main text (Section 5). The prediction-then-measurement reading of Section 5 holds on all three: from the published numbers the guarantee certifies $A_{\mathcal{F}} \geq 96.5$ (liver CT), 90.1 (brain MRI) and 95.4 (retina), and the checkpoints measure 96.3, 91.0 and 95.9: above the bound in every case, and within a point of it where $A_{\mathcal{B}}$ is at ceiling.

Table 20 repeats the comparison of Section 5 at $k \in \{4, 8, 16\}$ for the 6 detectors run at every budget, on the same six medical benchmarks and through the same code path. Whether the memory is given four normal images or sixteen changes how good the detectors are; it does not change whether the two metrics order them the same way. That is what the identity predicts, since w is a property of the data and not of the detector.

Table 21 runs the same detectors on MVTec AD and VisA. The industrial supports are k=16 unlabelled normals per category, two draws, taken from MVTec's own train split; this copy of VisA

Table 19: What the reported metric tracks on liver CT. For each re-run arm, the map is thresholded at the organ's own area fraction and scored by Dice against the organ mask (the border-rule foreground), and at the lesion's area fraction against the lesion mask; mean over images and both support draws, with A and $A_{\mathcal{F}}$ from Table 18. Every arm whose A is above 90 draws the organ; the one arm whose A is 4.5 draws the background, at the same $A_{\mathcal{F}}$ as the rest.

Detector	A	$A_{\mathcal{F}}$	organ Dice	lesion Dice
PatchCore features (WRN50-2), k supports	95.3	50.7	69.2	19.4
AnomalyDINO recipe (DINOv2 S/14)	94.4	65.3	55.0	24.1
memory, DINOv3 S/16@224	97.0	60.9	76.9	24.5
memory, DINOv3 B/16@224, k_{nn}=3	97.0	61.5	76.8	25.5
memory, DINOv3 B/16@336	96.7	56.3	80.9	34.9
memory, DINOv3 H+/16@224	96.6	69.3	53.5	20.6
memory, B/16+H+/16 (reported)	97.1	64.7	75.6	34.8
memory, B/16+H+/16, no calibration	97.2	66.0	73.5	34.1
memory + background masked out (AnomalyDINO)	4.5	63.4	0.1	0.0
memory + query masking	93.3	67.6	79.4	40.2
memory + map smoothing (PatchCore)	97.2	64.6	77.2	36.9

Table 20: The re-ranking is not an artefact of the support-set size. Kendall's τ between the ranking under whole-image pixel AUROC and the ranking under $A_{\mathcal{F}}$, with the number of detector pairs that change order in brackets (out of 15), for the 6 detectors run at every shot count on the six medical benchmarks. The disagreement tracks w, not k: it is largest on liver CT (w=0.94) at every k and absent on breast ultrasound (w=0.15) at every k.

k	Liver CT	Retina	Brain MRI	Head CT	Brain tum.	Breast US
4	0.07 (7)	1.00 (0)	0.33 (5)	0.47 (4)	0.60 (3)	0.87 (1)
8	0.07 (7)	1.00 (0)	0.20 (6)	0.47 (4)	0.33 (5)	1.00 (0)
16	0.07 (7)	0.73 (2)	0.47 (4)	0.60 (3)	0.60 (3)	1.00 (0)

ships no split file, so its supports are drawn from a reserved pool of normals that is excluded from the test split for every detector alike. Two things are visible. The inflation is present on the objects, which have a separable background ($\bar{w}$=0.57): the model-free rows score 82.9 and 77.1 under A and fall to 63.2 and 48.8 under $A_{\mathcal{F}}$. But the detectors' own foreground discrimination is high here (the best averages $A_{\mathcal{F}}$=97.7), so the whole-image number is not hiding much for them and their two rankings mostly agree (median τ=0.86 on objects, 1.00 on textures, where the rule finds little background, $\bar{w}$=0.36). The worst disagreement is on `pill` (τ=0.44, w=0.61). This is the identity's prediction: the whole-image and foreground orderings can differ only between detectors that differ in $A_{\mathcal{B}} - A_{\mathcal{F}}$, and on these objects the strong memories all have $A_{\mathcal{B}} \approx A_{\mathcal{F}} \approx 1$. Where foreground discrimination is weak, as on liver CT for every detector we ran, the first term is what the reported number measures, and the order it induces is the order of organ finding, not lesion finding.

H THE REFERENCE BASELINE

Setup. Support sets hold $k \in \{4, 8, 16\}$ normals per class, drawn independently per (k, seed) with two seeds. Anomaly maps are evaluated at 256×256 following Guo et al. (2025a): image-level AUROC on all nine datasets, pixel-level metrics on the six with masks, on a fixed random sample of 10^7 pixels per class, with the tissue-restricted variants on the same sample intersected with the input thresholded at 0.1. AUROC is reported raw, never reflected about 0.5. Every cell trains the decoder for the same 3000 steps (AdamW, lr 2×10^{-4}, batch 16), so shot count is never confounded with training length. Every tunable choice (taps, backbone, blocks, resolution, neighbour count, the per-class reduction, the fusion weight) is selected on the *validation* splits these datasets ship (2122 images, disjoint from test) at k=8 seed 0, then held fixed; no test image takes part in any decision.

The trained channel and the fusion. The second channel is one transformer block unrolled for eight rounds with shared weights, masked attention and a dropout bottleneck, trained on the normal support images only, with S_{rec} the training discrepancy read off at inference (Appendix O). Before

Table 21: The re-ranking outside medical imaging: the same detectors, the same code path, on MVTec AD and VisA with k=16 unlabelled normals per category (two draws averaged). Columns give the mean whole-image pixel AUROC A and the mean foreground term $A_{\mathcal{F}}$ over the categories of each group. On the objects, which have a separable background and a large w, the inflation is present (the model-free rows) but the frozen-prior memories localise well inside the object, so their two columns move together; on the textures the estimator finds little background and there is nothing to inflate. Across all 27 categories the median Kendall τ between the two rankings is 0.89, and the detector that wins under A differs from the one that wins under $A_{\mathcal{F}}$ on 6 of them. Where a detector's foreground discrimination is itself weak, as for every detector on liver CT, the two columns diverge; where it is not, they cannot, which is what Eq. 1 says.

	MVTec objects (10)		MVTec textures (5)		VisA (12)	
Detector	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$	A	$A_{\mathcal{F}}$
model-free predictor (no model)	77.6	56.2	67.9	58.5	87.4	69.0
foreground mask (no model)	73.6	48.6	58.8	40.6	80.0	49.0
PatchCore features (WRN50-2), k supports	95.5	94.7	94.8	94.5	92.4	94.7
AnomalyDINO recipe (DINOv2 S/14)	95.7	95.1	96.8	96.5	96.7	96.7
memory, DINOv3 B/16@336	97.4	96.0	98.2	97.9	98.8	97.6
memory, DINOv3 H+/16@224	97.2	95.8	97.7	97.3	98.3	97.5
memory, B/16+H+/16 (reported)	98.0	96.9	98.4	98.1	99.2	98.4
memory + background masked out (AnomalyDINO)	83.8	96.2	98.4	98.1	92.1	98.1
AnomalyCLIP (Zhou et al., 2024), released (cross-dataset zero-shot)	88.0	79.6	97.2	96.9	95.9	89.7
median Kendall τ	0.86		1.00		0.86	
mean w	0.50		0.36		0.62	

Table 22: One unified model over nine datasets and six imaging modalities, from k unlabelled normal images per class and no training; mean ± std over two independent support sets. The second block is the metric the field reports; the third restricts it to tissue pixels.

k	Brain MRI	Liver CT	Retina RESC	Retina OCT17	Chest X-ray	Histopath.	Breast US	Head CT	Brain tumour	**Avg.**
Image-level AUROC (%, ↑)										
4	80.4 ± 2.2	61.4 ± 0.0	93.0 ± 0.7	90.4 ± 2.4	81.5 ± 0.6	64.5 ± 6.1	83.5 ± 4.1	85.0 ± 3.6	98.1 ± 1.6	**82.0 ± 0.3**
8	84.1 ± 0.4	64.2 ± 2.7	94.2 ± 2.0	97.2 ± 0.5	84.5 ± 1.1	72.8 ± 2.2	90.2 ± 0.6	92.4 ± 0.8	99.1 ± 0.5	**86.5 ± 0.4**
16	88.5 ± 0.5	64.9 ± 1.4	94.6 ± 0.8	97.5 ± 0.0	84.5 ± 0.2	73.4 ± 1.0	91.2 ± 0.3	96.5 ± 0.8	99.6 ± 0.1	**87.8 ± 0.0**
Pixel AUROC, whole image (%, ↑)										
4	95.9 ± 0.5	96.8 ± 0.2	97.1 ± 0.2	–	–	–	79.7 ± 1.7	94.1 ± 0.9	92.8 ± 0.7	**92.7 ± 0.6**
8	96.7 ± 0.1	96.9 ± 0.3	97.2 ± 0.5	–	–	–	80.7 ± 0.0	96.3 ± 0.1	93.4 ± 0.1	**93.5 ± 0.0**
16	97.0 ± 0.0	97.1 ± 0.0	97.6 ± 0.1	–	–	–	81.0 ± 0.1	96.2 ± 0.3	94.1 ± 0.6	**93.8 ± 0.1**
Pixel AP, tissue only (%, ↑)										
4	19.1 ± 1.0	10.5 ± 0.9	69.7 ± 2.0	–	–	–	33.0 ± 0.6	42.8 ± 5.1	45.6 ± 3.0	**36.8 ± 1.8**
8	22.7 ± 0.9	12.1 ± 0.8	72.1 ± 4.3	–	–	–	34.2 ± 0.0	56.1 ± 0.4	46.8 ± 0.7	**40.7 ± 0.6**
16	24.2 ± 0.1	13.8 ± 1.0	75.3 ± 1.1	–	–	–	35.3 ± 0.8	54.8 ± 2.2	48.2 ± 3.0	**41.9 ± 0.7**

fusing, each channel is standardised within each class, $z(\cdot) = (\cdot - \mu)/\sigma$, and combined as $S = (1 - w_f)\, z(S_{\text{rec}}) + w_f\, z(S_{\text{ret}})$ with one scalar w_f shared by all classes and shot counts. The statistics are computed over the test maps of a class, which makes the calibration transductive; it is affine on a single map, and on the reported model, which has no decoder, removing it moves image AUROC by at most 0.05 points (Appendix Q).

Results. Averaged over the nine datasets, image-level AUROC rises from 82.0 at k=4 to 86.5 at k=8 and 87.8 at k=16; whole-image pixel AUROC is 93.5 at k=8 and the tissue-restricted pixel AP 40.7. The seed-to-seed spread is under a point on every image-level aggregate. Retinal OCT is where the memory works (94.2 image AUROC at k=8, tissue-restricted pixel AP 0.72 against a lesion prior of 0.059); liver CT and histopathology are weak (64.9 and 73.4 at k=16): a frozen natural-image prior transfers to retinal layers and brain anatomy far better than to liver parenchyma or stained tissue. On the development cell (k=8, seed 0, both channels at ViT-S/16) the decoder reaches 55.7 image AUROC after 3000 gradient steps on the same k images and the memory 69.1 with nothing trained; no intermediate fusion weight beats the memory alone (Appendix K). The decoder is not useless, only redundant: at a 5% false-alarm budget it catches 93 brain MRI and 208 chest X-ray anomalies the memory misses (Appendix L), and a single global weight cannot exploit that.

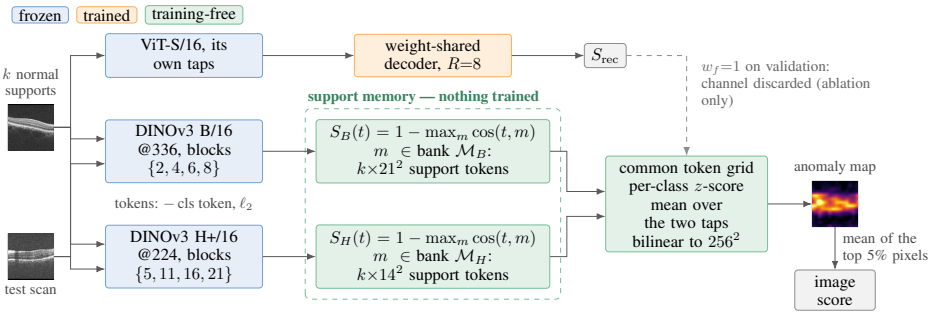


Figure 7: The two channels of the baseline. The k normal support images are encoded by frozen DINOv3 taps and used twice: as a nearest-neighbour token memory (bottom, no training, where the performance comes from) and as targets for the weight-shared adapter decoder (top). The memory runs over two taps, B/16@336 + H+/16@224, and the decoder over its own single ViT-S/16. Each tap keeps its own support bank and produces its own map; the two maps are brought to a common grid, standardised per class, averaged and upsampled to 256^2, and the reduction from map to image score is selected per class on validation. The selected fusion weight discards the top channel.

Table 23: Image-level AUROC (%) on the BMAD/MVFA six, grouped by what each method is given. *Normals* is how many normal images the row sees; the best is marked within each budget, since comparing across budgets is the category error this table exists to avoid. Upper block: normal images and no anomaly labels, our supervision class, at two budgets; MKD and PatchCore see the entire normal training set, so the matched row from us is the one that also sees it, retrieving its 16 references per query. The AnomalyDINO row is *our* run of that recipe on these splits, since it publishes industrial numbers only. Lower block: k *annotated* support samples and one model per dataset; those methods are ahead of us and the table says so. Published values are transcribed from MVFA's or MadCLIP's Table 1. Ours is one model shared by all nine datasets of Table 22.

Method	Year	Normals	Liver CT	RESC	Brain	OCT17	Chest	Histo.	Avg.
normal images, no anomaly labels — our supervision class									
MKD (Salehi et al., 2021)	2021	all	60.4	89.0	81.4	96.6	82.0	77.7	81.2
PatchCore (Roth et al., 2022)	2022	all	60.4	91.5	91.5	98.6	75.2	69.3	81.1
SUMMA (ours), pool-retrieved	–	all	63.6	95.4	92.1	99.7	87.0	76.2	**85.7**
AnomalyDINO (Damm et al., 2025) recipe *(our run)*	2025	16	64.1	85.0	85.0	81.6	76.9	67.8	76.7
SUMMA (ours)	–	16	64.9	94.6	88.5	97.5	84.5	73.4	**83.9**
k=16 *annotated support samples, one model per dataset*									
APRIL-GAN (Chen et al., 2023)	2023	16	82.9	96.0	94.0	99.9	78.6	81.2	88.8
MediCLIP (Zhang et al., 2024)	2024	16	79.3	86.5	91.6	96.4	69.7	70.2	82.3
MVFA (Huang et al., 2024)	2024	16	83.8	97.2	94.4	99.7	85.7	82.6	90.6
MadCLIP (Shiri et al., 2025)	2025	16	91.5	99.1	95.9	99.7	88.2	90.1	**94.1**

I Comparison, per dataset

Table 23 groups published results by what they are given, and within our class by how many normals. Table 24 gives the three datasets every published method reports per dataset and shot count. The same baseline is reported differently by different papers (MVFA's own k=2 histopathology figure is 82.61, while a later paper reproducing it lists 62.05), so we label each row with the paper it was transcribed from and do not mix sources within a row; the image-level and pixel-level columns of this literature are also easy to confuse, which is part of why we report both.

I.1 A fixed budget against a retrievable pool

The second block of Table 24 is training-free and label-free like ours, yet ahead of us, and the reason is not the mechanism: that method keeps the entire normal training pool at inference and retrieves its K references *per query* (Huo et al., 2026). Its "K=4" is four images chosen from thousands, per scan. We measured that difference rather than asserting it.

Table 24: Image-level AUROC (%) on the three datasets every published few-shot paper reports, at each shot count. The first block consumes k *annotated* support samples and fits one model per dataset. The second block is training-free and label-free like ours but is *not* few-shot: its K counts references retrieved *per query* from the entire normal pool (1,542 liver, 4,297 retina, 7,500 brain images), so it is placed in the k=4 columns only because that is the K its paper reports, and it should be read as a full-pool method. Ours uses k unlabelled normals, full stop, and one model for nine datasets. Values are transcribed from the paper credited in the row — the same baseline is reported differently by different papers, so a row is comparable only within its source.

	k=4			k=8			k=16		
Method	Liver	RESC	Brain	Liver	RESC	Brain	Liver	RESC	Brain
k annotated support samples, one model per dataset									
BGAD (Yao et al., 2023)	72.48	86.22	83.56	74.60	89.96	88.01	78.79	91.29	88.05
MediCLIP (Zhang et al., 2024)	76.48	89.37	88.43	82.04	88.30	90.77	84.11	87.67	92.03
APRIL-GAN (Chen et al., 2023)	53.05	94.70	89.18	62.38	91.36	88.41	82.94	95.96	94.03
MVFA (Huang et al., 2024)	81.18	96.18	92.44	85.90	96.57	92.61	83.85	97.25	94.40
MadCLIP (Shiri et al., 2025)	82.97	96.62	95.95	89.31	97.16	95.17	91.46	99.11	95.90
Spatial-FAD (Miao et al., 2026)	92.39	94.72	93.93	94.69	97.23	91.45	95.59	96.83	95.43
training-free, no labels, but the whole *normal pool available at inference*									
HiMatch-AD (Huo et al., 2026)	77.84	92.61	92.91	–	–	–	–	–	–
SUMMA (ours)	61.39	93.03	80.42	64.22	94.23	84.14	64.90	94.60	88.48

Table 25: What the fixed budget costs. Image-level AUROC (%), same taps, same calibration and the same per-class aggregation in every column; the *only* difference is whether the k normal images were drawn once in advance (FIXED, the setting this paper argues for) or retrieved per test image from the whole normal training pool (RETR., the setting concurrent training-free work runs in). Datasets are ordered by pool size, which is the hypothesis this table was built to test and which it refutes: the last three hold fewer than a hundred normals and gain *more* at k=4 than the six whose pools run to thousands. What the advantage tracks is k, not the pool. This is a measurement of the retrieval policy under our own method, not a reproduction of any published system.

		k=4			k=16		
Dataset	pool	fixed	retr.	Δ	fixed	retr.	Δ
Retina OCT17	26,315	92.8	99.4	**+6.6**	97.4	99.7	**+2.2**
Chest X-ray	8,000	82.1	85.5	**+3.4**	84.4	87.0	**+2.6**
Brain MRI	7,500	78.2	89.5	**+11.4**	88.9	92.1	**+3.2**
Histopath.	5,088	70.6	75.7	**+5.1**	74.4	76.2	+1.8
Retina RESC	4,297	92.4	95.3	**+2.9**	93.8	95.4	+1.6
Liver CT	1,542	61.4	62.8	+1.4	63.5	63.6	+0.1
Breast US	99	79.4	93.5	**+14.1**	90.8	94.2	**+3.4**
Head CT	90	81.4	97.3	**+15.9**	97.4	99.4	**+2.0**
Brain tumour	88	96.5	98.6	**+2.1**	99.7	99.6	-0.1
Mean		**81.6**	**88.6**	**+7.0**	**87.8**	**89.7**	**+1.9**

Table 25 runs both policies through one code path on all nine datasets, the only difference being whether the k images are drawn once in advance or retrieved for each test image from that class's whole normal pool. Retrieval is worth 7.0 points at k=4 and 1.9 at k=16: choosing the support set matters most when the budget is too small to cover the normal appearances by luck. It is not bought with the pool: the 3 datasets whose pools hold at most 99 images gain 10.7 points at k=4, *more* than the 6 whose pools run to 26,315, which gain 5.1. And it does not explain the concurrent result: on the three datasets of Huo et al. (2026) our retrieved arm reaches 82.5 at k=4 against their 87.8, the shortfall sitting almost entirely on liver CT (62.8 against their 77.8), a difference of method, not of budget.

Table 26: One unified SUMMA model over nine datasets and six imaging modalities, from k unlabelled normal images per class and no training. Mean ± std over two independent support sets. The second block is the metric the field reports; the third restricts it to tissue pixels, which is the one we argue for.

k	Brain MRI	Liver CT	Retina RESC	Retina OCT17	Chest X-ray	Histopath.	Breast US	Head CT	Brain tumour	**Avg.**
					Image-level AUROC (%, ↑)					
1	65.0 ± 0.1	60.2 ± 0.1	84.7 ± 7.3	96.2 ± 0.2	62.9 ± 19.2	50.8 ± 9.9	81.2 ± 1.3	72.2 ± 14.3	98.7 ± 0.4	**74.7 ± 5.5**
2	80.6 ± 1.5	64.8 ± 0.7	91.2 ± 0.9	89.4 ± 2.8	83.2 ± 0.8	67.6 ± 6.4	87.0 ± 0.5	86.9 ± 2.5	97.6 ± 1.6	**83.2 ± 1.1**
4	80.4 ± 2.2	61.4 ± 0.0	93.0 ± 0.7	90.4 ± 2.4	81.5 ± 0.6	64.5 ± 6.1	83.5 ± 4.1	85.0 ± 3.6	98.1 ± 1.6	**82.0 ± 0.3**
8	84.1 ± 0.4	64.2 ± 2.7	94.2 ± 2.0	97.2 ± 0.5	84.5 ± 1.1	72.8 ± 2.2	90.2 ± 0.6	92.4 ± 0.8	99.1 ± 0.5	**86.5 ± 0.4**
16	88.5 ± 0.5	64.9 ± 1.4	94.6 ± 0.8	97.5 ± 0.0	84.5 ± 0.2	73.4 ± 1.0	91.2 ± 0.3	96.5 ± 0.8	99.6 ± 0.1	**87.8 ± 0.0**
32	89.9 ± 0.7	64.3 ± 0.6	95.4 ± 0.4	97.9 ± 0.1	85.5 ± 0.0	75.2 ± 0.0	90.0 ± 1.1	99.1 ± 0.5	99.8 ± 0.0	**88.6 ± 0.1**
					Pixel AUROC, whole image (%, ↑)					
1	94.8 ± 0.5	96.1 ± 0.4	96.3 ± 1.0	–	–	–	77.5 ± 1.7	91.6 ± 0.2	91.2 ± 1.7	**91.3 ± 0.7**
2	96.0 ± 0.1	96.5 ± 0.1	97.0 ± 0.1	–	–	–	78.6 ± 0.9	93.1 ± 2.2	91.2 ± 1.2	**92.1 ± 0.7**
4	95.9 ± 0.5	96.8 ± 0.2	97.1 ± 0.2	–	–	–	79.7 ± 1.7	94.1 ± 0.9	92.8 ± 0.7	**92.7 ± 0.6**
8	96.7 ± 0.1	96.9 ± 0.3	97.2 ± 0.5	–	–	–	80.7 ± 0.0	96.3 ± 0.1	93.4 ± 0.1	**93.5 ± 0.0**
16	97.0 ± 0.0	97.1 ± 0.0	97.6 ± 0.1	–	–	–	81.0 ± 0.1	96.2 ± 0.3	94.1 ± 0.6	**93.8 ± 0.1**
32	97.1 ± 0.1	97.3 ± 0.1	97.7 ± 0.2	–	–	–	82.2 ± 0.5	96.4 ± 0.0	94.2 ± 0.0	**94.2 ± 0.1**
					Pixel AP, tissue only (%, ↑)					
1	15.5 ± 0.7	7.1 ± 0.2	60.5 ± 8.9	–	–	–	27.5 ± 1.0	31.6 ± 1.4	42.6 ± 5.5	**30.8 ± 2.3**
2	20.9 ± 1.2	9.6 ± 0.9	69.8 ± 0.4	–	–	–	31.2 ± 0.7	39.4 ± 9.0	40.2 ± 3.3	**35.2 ± 2.1**
4	19.1 ± 1.0	10.5 ± 0.9	69.7 ± 2.0	–	–	–	33.0 ± 0.6	42.8 ± 5.1	45.6 ± 3.0	**36.8 ± 1.8**
8	22.7 ± 0.9	12.1 ± 0.8	72.1 ± 4.3	–	–	–	34.2 ± 0.0	56.1 ± 0.4	46.8 ± 0.7	**40.7 ± 0.6**
16	24.2 ± 0.1	13.8 ± 1.0	75.3 ± 1.1	–	–	–	35.3 ± 0.8	54.8 ± 2.2	48.2 ± 3.0	**41.9 ± 0.7**
32	24.3 ± 0.9	15.4 ± 1.2	76.4 ± 1.4	–	–	–	36.8 ± 0.8	60.6 ± 0.7	48.3 ± 0.6	**43.6 ± 0.5**

Table 27: Sensitivity to the retrieval fusion weight w on the development cell (k=8, seed 0), with and without per-class calibration. w=0 is reconstruction only.

Calibration	w=0	w=0.1	w=0.2	w=0.3	w=0.5	w=0.7	w=1
none	55.7	55.7	55.7	55.7	55.8	55.8	69.1
per-class	55.7	57.5	59.6	62.2	66.9	68.5	69.1

J Complete per-class results and the full shot curve

The spread is not uniform: from $k \geq 4$ brain MRI and chest X-ray are stable (within 2.6 points between seeds), while histopathology moves by 8.7 points between the two k=4 support sets and retinal OCT by 3.9 at k=8. At k=1 nothing is stable: chest X-ray moves by 24.1 points. A single support set is not a sufficient experiment in this regime.

Table 26 reports k from 1 to 32. k=1 is not a working regime: the mean image AUROC is 74.7, but the standard deviation between the two support sets is 5.5 points overall and 12.1 on chest X-ray; by k=8 the spread is 0.2 points. The curve is not monotone at the low end (k=2 scores 83.2 and k=4 scores 82.0, within noise). One dataset stops improving and the rest do not: brain MRI climbs from 65.0 to 89.9 and retinal RESC from 84.7 to 95.4 as k goes from 1 to 32, while liver CT never leaves the low sixties. On liver the bottleneck is the representation, not support-set coverage.

K Fusion sweep

Everything in this section is on the five-dataset subset the decoder was trained on, at ViT-S/16; it is the ablation, not the reported model (the per-cell ablation and the full sweep are in the supplementary results files). The scan of the fusion weight returned w_f=1, the degenerate case: all the weight on the support memory and none on the decoder. Table 27 shows why. Without calibration the curve is flat at 55.7–55.8 for every $w_f < 1$ and then jumps to 69.1 at w_f=1: the decoder's raw scores are on a large enough scale that any admixture of it swamps the memory entirely. With per-class calibration the curve is monotone, $55.7 \to 57.5 \to 59.6 \to 62.2 \to 66.9 \to 68.5 \to 69.1$: every step towards the memory helps, the shape of a weighted average between a weak signal and a strong one, not of two complementary ones.

Table 28: Training-free design sweep of the support memory on the **validation** split, with the k=8 seed-0 support set. Image AUROC is averaged over five datasets, the pixel columns over the three with masks. Best in each column in bold.

Backbone	Blocks	Res.	k_{nn}	Image AUROC			Pixel	Pixel AP
				max-pool	best	agg.	AUROC	(tissue)
ViT-B/16	$\{2,4,6,8\}$	224	1	70.4	72.8	top5	96.5	34.1
ViT-B/16	$\{2,4,6,8\}$	224	3	70.2	73.0	top5	96.5	**34.2**
ViT-B/16	$\{2,4,6,8\}$	336	1	71.2	73.7	top5	96.5	33.9
ViT-B/16	$\{2,4,6,8\}$	336	3	70.7	73.6	top5	96.5	33.6
ViT-B/16	$\{2,4,6,8\}$	448	1	70.7	74.5	top5	96.5	32.8
ViT-B/16	$\{2,4,6,8\}$	448	3	70.6	74.4	top5	96.4	32.4
ViT-S/16	$\{2,4,6,8\}$	224	1	67.9	70.8	top5	96.1	31.4
ViT-S/16	$\{2,4,6,8\}$	224	3	68.5	70.8	top5	96.2	31.2
ViT-S/16	$\{2,4,6,8\}$	336	1	68.1	70.5	top5	96.1	30.2
ViT-S/16	$\{2,4,6,8\}$	336	3	67.1	70.6	top5	96.1	29.8
ViT-S/16	$\{2,4,6,8\}$	448	1	68.3	70.3	top5	95.8	29.3
ViT-S/16	$\{2,4,6,8\}$	448	3	67.5	70.3	top5	95.7	28.7
ViT-S/16	$\{4,6,8,10\}$	224	1	71.8	73.1	top5	96.3	32.0
ViT-S/16	$\{4,6,8,10\}$	224	3	71.1	73.1	top5	96.4	32.0
ViT-S/16	$\{4,6,8,10\}$	336	1	70.7	74.0	top5	96.5	32.6
ViT-S/16	$\{4,6,8,10\}$	336	3	70.5	73.8	top5	96.5	32.4
ViT-S/16	$\{4,6,8,10\}$	448	1	72.5	**75.0**	top5	96.4	32.6
ViT-S/16	$\{4,6,8,10\}$	448	3	**72.6**	74.8	top5	96.4	32.3

L Failure cases and qualitative results

At a 5% false-alarm budget (the 95th percentile of each class's *normal* scores, placed without reference to any anomaly label) the reported configuration captures 54.6% of brain MRI anomalies, 20.6% of liver CT, 25.0% of chest X-ray, 3.9% of histopathology and 74.9% of retinal OCT at k=16 (Table 32); AUROC hides how sobering that is. In the ablation cell, where a decoder channel exists to compare against, the decoder catches 93 brain MRI and 208 chest X-ray anomalies that the memory misses at the same budget; a per-class or per-image gate is the obvious next thing to try, and we did not try it.

The retrieval channel's false positives are, by construction, normal appearances absent from the k support images; the reconstruction channel fails the other way round, reconstructing an *anomalous* region whose local texture resembles something in the support set.

M Design sweep for the support memory

Because the memory needs no training, its configuration can be searched exhaustively in minutes. The sweep covers input resolution, backbone scale, which blocks are tapped, and how many neighbours are averaged, every entry using the same k=8 seed-0 support set. Table 28 reports it on the validation splits, where the choice was made; the same 18-cell grid repeated on a subsample of the test split (supplementary results files) differs by 0.3 criterion points at its optimum.

Four things are visible. The image-score reduction matters more than resolution or k_{nn}: the mean of the top 5% of pixels gains over the maximum of a 16×16 average-pooled map in every one of the 36 cells, by 1.3 to 6.0 points of image AUROC, while resolution and neighbour count together move it by at most 1.2. Scale helps, but tap depth does much of the work: moving ViT-S/16's taps from $\{2,4,6,8\}$ to $\{4,6,8,10\}$ closes most of the gap to ViT-B/16 at no cost. Resolution does not help: for every backbone, 448px gives the lowest tissue-restricted AP of the three resolutions. More neighbours do not help. The transfer is imperfect: validation selects ViT-B/16 at 336px, the test sweep's own optimum is ViT-B/16 at 224px.

Table 29: The frozen-memory literature's standard additions, at k=16, seed 0. CRIT is the selection criterion declared in the paper, the mean of image AUROC and tissue-restricted AP. Two things to read here. Masking the memory — the trick AnomalyDINO reports as helping across the board industrially — **collapses whole-image pixel AUROC while leaving the tissue-restricted score alone**, because on these images the tokens it removes are what that metric was scoring. And fg-query wins on validation and then largely does not transfer, so the reported configuration is unchanged: adopting it would have traded 2.9 points of the metric this paper criticises for 0.6 of the metric it advocates, chosen after seeing test.

		pixel AUROC			
Variant	image	whole	tissue	tis. AP	CRIT
validation — where the choice was made					
none (reported)	79.8	97.3	85.3	37.7	58.8
fg-mem *(AnomalyDINO's masking)*	70.1	44.7	85.0	36.3	53.2
fg-query	80.3	96.7	86.6	39.5	59.9
fg-mem + fg-query	78.2	95.0	86.7	39.1	58.7
smooth σ=4 *(PatchCore's)*	79.8	97.4	85.2	37.8	58.8
test — reported for transparency, not used to choose					
none (reported)	86.3	94.0	86.9	42.6	64.4
fg-query	86.5	91.0	86.6	43.2	64.9

M.1 What the memory literature adds, and what it does here

The patch-memory line masks background tokens out of the support memory (Damm et al., 2025) and smooths the anomaly map (Roth et al., 2022). We added both, plus the query-side counterpart of the masking, using the paper's own model-free tissue threshold rather than a segmenter, and swept them on validation with the reported configuration (Table 29). Masking the memory does the opposite of what it does industrially, and Section 2 says why: it takes whole-image pixel AUROC from 97.3 to 44.7, a fall of 52.6 points, while the tissue-restricted score barely moves (85.3 to 85.0). With no background tokens in the memory, background query tokens have nothing to match and score as maximally anomalous, so $A_{\mathcal{B}}$ collapses and the term that was carrying w=0.94 of the metric on liver CT stops being fed; $A_{\mathcal{F}}$ is unaffected. Query masking leads the criterion by 1.15 points on validation and by 0.41 on test, where the tissue-restricted AUROC gain vanishes (-0.3) and what survives is $+0.6$ of tissue AP bought with -2.9 of whole-image pixel AUROC; we report the configuration unchanged. Smoothing is a wash either way.

M.2 Scaling the prior, and why two taps beat one

A larger prior costs only inference, so we also tried DINOv3 ViT-H+/16 (840M parameters, 32 blocks, 1280-d tokens). The taps must scale with depth: blocks $\{2, 4, 6, 8\}$ are far too shallow for a 32-block model and cost 14 points of image AUROC in the tap sweep. And 336px does not fit in 24 GB, so ViT-H+ could only be compared at 224px, on blocks $\{5, 11, 16, 21\}$. On the declared criterion ViT-H+ finishes ahead of ViT-B on validation by only 0.4 points (54.7 against 54.3). Table 30 resolves that on the full test split: ViT-H+ is ahead on every aggregate at every k, by 3–5 points of tissue-restricted AP (34.4 against 31.2 at k=8), but per dataset it gains 1–4 points on brain MRI, retinal OCT, chest X-ray and histopathology and *loses* 2–6 on liver CT, the dataset dominated by the organ mask.

That per-dataset split is what makes the ensemble work rather than merely average. Table 31 reports the selection, and it contains its own control: running *one* prior at two resolutions and averaging already gains (55.2 against 54.3), so part of the effect is variance reduction across two correlated maps; two *different* priors gain more (55.9) and win every column rather than trading one against another.

Table 30: Backbone scale on the *full* test split, averaged over the five modalities and two support sets. The two single-tap rows are what the ensemble is built from; ViT-H+/16 is ten times larger than ViT-B/16 and leads it on every aggregate here, though not on every dataset (it loses 2–6 points on liver CT). The ensemble row is the configuration the main paper reports. Blocks are $\{2, 4, 6, 8\}$ for B/16 and $\{5, 11, 16, 21\}$ for H+/16; subscript t marks the tissue-restricted variants.

Tap(s)	k	Img. AUC	Pix. AUC	Pix. AUC$_t$	Pix. AP$_t$
B/16@336	4	73.0	95.7	76.9	28.4
	8	76.4	96.2	79.9	31.2
	16	77.6	96.5	79.7	32.6
H+/16@224	4	74.8	96.0	81.5	32.0
	8	76.7	96.4	83.2	34.4
	16	78.5	96.8	84.7	37.6
B/16@336 + H+/16@224	4	82.0	92.7	84.6	36.8
	8	86.5	93.5	86.5	40.7
	16	87.8	93.8	86.9	41.9

Table 31: Choosing the prior ensemble on the **validation** split (k=8, seed 0). The criterion in the last column — the mean of image AUROC and tissue-restricted pixel AP — was fixed before the search. Row three is the control: averaging one prior's maps at two resolutions already gains, so part of the effect is variance reduction; two *different* priors gain more than that, and beat both single priors in every column. The last two rows are separated by 0.003 criterion points, far inside the seed-to-seed spread of the main results, so we take the cheaper of the two — a tie-break on cost, which here selects the marginally *lower*-scoring row. Selected row in bold.

Prior(s)	Img. AUC	Pix. AUC	Pix. AUC$_t$	Pix. AP$_t$	Crit.
B/16@336	74.7	96.5	81.6	33.9	54.3
H+/16@224	73.6	96.3	83.4	35.7	54.7
B/16@224 + B/16@336	74.4	96.8	82.6	36.0	55.2
B/16@336 + H+/16@224	74.9	97.0	83.8	37.0	55.9
B/16@224 + B/16@336 + H+/16@224	74.5	97.0	83.7	37.4	55.9

N Operating points

AUROC is threshold-free, which is convenient for comparison and useless for deployment, so Table 32 reports the fraction of anomalies caught at 1% and 5% false-alarm rates and the false-alarm rate needed to catch 95%.

O Implementation details

Support memory. Two frozen DINOv3 taps, selected on validation: ViT-B/16 at 336×336 read at blocks $\{2, 4, 6, 8\}$ ($4 \times 768 = 3072$ channels, 21×21 tokens) and ViT-H+/16 at 224×224 read at blocks $\{5, 11, 16, 21\}$ ($4 \times 1280 = 5120$ channels, 14×14 tokens), tokens context-recentred by subtracting the block's class token (Guo et al., 2025b) and ℓ_2-normalised. Each tap keeps its own bank (441k vectors of 3072 dimensions and 196k of 5120, together 151 MB at fp32 for k=16) and produces its own map; the two maps are brought to the coarser grid, standardised per class and averaged. Scoring one image costs $2N^2kC$ FLOPs per tap, 1.6–25 GFLOPs over the two taps for k=1–16.

Decoder. ViT-S/16 with the same taps ($4 \times 384 = 1536$ channels). Hidden width 1024, 8 unrolled rounds with shared attention and adapter weights and per-round LayerNorms, 8 heads, 9×9 neighbourhood-masked attention, adapter bottleneck 512 with zero-initialised up-projection and learnable γ, dropout bottleneck p=0.4. Four loose-reconstruction groups at rounds $\{2, 4, 6, 8\}$, each regressing a fixed random orthonormal $1536 \to 256$ projection of its encoder tap, hard-mining ratio 0.25, auxiliary weight 0.5. AdamW, lr 2×10^{-4}, $\beta = (0.9, 0.999)$, weight decay 10^{-4}, gradient

Table 32: What the reported image score is worth at a threshold someone would actually set, for the configuration the paper reports, averaged over two support sets. cap@far$q$: the fraction of anomalies caught while holding the false-alarm rate on that class's normals to q. far@cap95%: the false-alarm rate needed to catch 95% of anomalies. The same image score produces the AUROCs in Table 26, and the contrast between the two is the point. These are read off the top-5% score throughout, including for the classes whose reported AUROC uses a different per-class reduction, so for those rows they are a lower bound.

k	metric	Brain MRI	Liver CT	Retina RESC	Retina OCT17	Chest X-ray	Histopath.	Breast US	Head CT	Brain tumour
4	cap@far5%	36.8	17.0	70.0	66.2	27.5	4.6	31.5	57.5	92.9
	cap@far1%	6.7	4.5	51.4	50.3	4.7	1.6	21.4	56.0	91.9
	far@cap95%	59.8	77.1	35.4	43.0	74.5	78.7	50.0	60.0	10.0
8	cap@far5%	39.2	19.5	76.6	86.4	27.8	3.0	42.0	65.5	97.7
	cap@far1%	8.0	9.7	62.0	68.1	5.9	0.9	18.6	63.0	97.4
	far@cap95%	62.7	81.8	34.9	19.8	67.3	69.0	39.7	20.0	0.0
16	cap@far5%	54.6	20.6	74.9	88.1	25.0	3.9	74.0	88.0	99.0
	cap@far1%	7.6	6.2	58.4	70.0	6.2	0.8	30.9	84.0	99.0
	far@cap95%	54.8	76.1	31.5	17.8	67.2	69.5	44.1	25.0	0.0

clipping 0.1, batch 16, 3000 optimiser steps per cell. 14.99 M trainable parameters on a 21.60 M frozen ViT-S/16, 45.88 GFLOPs per forward pass at 224×224. A single RTX 3090.

Evaluation. Maps are evaluated at 256×256 after Gaussian smoothing with σ=4. Pixel AUROC uses a fixed random sample of 10^7 pixels per class, drawn with replacement and seeded per class so that every arm of every sweep sees identical pixels.

The model-free audit. Images are converted to grey and resized to 256×256; the blur is a Gaussian with σ=4 truncated at half a standard deviation; the border estimator marks as background every pixel connected to the frame edge whose intensity is within 0.06 of the median of the outer four pixel rows and columns; the mask predictor is that estimated foreground. Scores are accumulated in 4096-bin histograms for anomalous, background-normal and foreground-normal pixels over every image of the split; AUROC, AP and F1-max follow exactly up to the bin width, and the per-image AUROC is computed per image. The identity's residual is at most $1e-13$ AUROC points on any row.

P Evaluation-pipeline corrections

Three defects in the evaluation code we inherited would each have silently corrupted the numbers in this paper, and all three are invisible on the industrial benchmarks the code was written for. **AUROC reflection**: the scoring helper reflected any AUROC below 0.5 to $1 - \mathrm{AUROC}$; on industrial texture no class scores below chance, so the branch never fires, but in few-shot medical AD a hard class at k=4 can and does, and the reflection turns a failure into an apparent success. **Score-map filename collisions**: per-image maps were keyed by class, parent directory and stem, which assumes the parent directory separates normal from anomalous images; here both live under the same directory and histopathology names its files `0.png`, `1.png`, ... in both, so 917 of 12,010 maps overwrote one another. **Placeholder masks**: a detection-only dataset yields an all-ones mask for every anomaly, and averaging its "pixel AUROC" into a benchmark mean produces a number that moves with nothing.

Q Limitations of the baseline

The calibration is transductive, and on the reported model not load-bearing. Per-class standardisation uses the statistics of that class's test maps. Applied to a single map it is affine and cannot change a rank-based metric; it acts only where maps are combined, and since the selected fusion weight drops the decoder, the only combination left is the average of the two frozen taps. Table 33 removes it from exactly that average: image AUROC moves by at most 0.05 points, and every pixel column is *better* without it. The reported model can be run per-image with no test-set statistics at all.

Table 33: What the transductive calibration is worth on the reported model. Both arms come from one run, so the encodings are identical and the only difference is whether each tap's map is standardised over the class before the two are averaged. Nine datasets, seed 0. The calibration is not load-bearing here: it costs at most 0.05 points of image AUROC, and every pixel column is *better* without it. It matters for the channel fusion (Table 27, 55.8 against 66.9) because a trained decoder's residual and a cosine distance are on different scales; two cosine distances are not.

	k=4		k=8		k=16	
Metric	cal.	none	cal.	none	cal.	none
image AUROC	80.8	81.1	85.1	85.0	86.3	86.2
pixel AUROC, whole image	92.1	92.3	93.6	93.7	94.0	94.1
pixel AUROC, tissue only	84.0	84.4	86.2	86.5	86.9	87.3
pixel AP, tissue only	35.0	35.4	41.3	41.8	42.6	43.2

The retrieval channel is bounded by support-set coverage. A healthy variant the support set omits is anomalous to it by construction.

Only training-free methods were re-run under our protocol. The other rows of the comparison table transcribe published numbers; those methods use per-dataset models, annotated support samples and in some cases different test-split conventions, so the comparison bounds where the field sits rather than settling a ranking.

Decoder capacity is not tuned for the few-shot regime, the token grid caps localisation at 21×21, ViT-H+/16 could not be run at 336px in 24 GB, and chest X-ray is subsampled, so its absolute number is not directly comparable to one computed on the full split.