ONE SCORE TERM TOO MANY: THE APPARENT COST OF NORMALIZING FROZEN VIT TOKENS IS LARGELY AN ANOMALY-SCORE ARTIFACT

Anonymous authors
Paper under double-blind review

ABSTRACT

Detectors that reconstruct frozen vision-transformer tokens usually ℓ_2-normalize each patch token first, and on periodic manufacturing texture this convention appears to be expensive: in a strictly paired counterfactual (one recipe, one boolean flipped) normalization costs +15.38, +38.64 and +23.04 mean anomaly-detection points (mAD) on wafer, SEM-circuit and PCB benchmarks. We show that this effect is largely a property of the *anomaly score*, not of the features or the objective. The family's score adds a scale-sensitive ℓ_2 residual to a cosine one, so normalizing or rescaling the reconstruction target reshapes the anomaly map. Rescoring the same checkpoints with the cosine residual alone, without retraining, removes the effect on all six datasets: on wafer the normalized arm moves from 49.39 to 66.77 mAD, level with the un-normalized arm (65.87), and a fixed-constant rescale that appeared to cost -12.16 mAD costs -0.04. A squared-error loss term also carries scale, by 2–4 mAD at every weight tried, with or without a cosine term beside it. Two public cosine-scored detectors are immune to the identical rescale; one of them, scored with an ℓ_2 residual instead, loses 25 mAD over three seeds, and the published UniAD checkpoint, scored with ℓ_2, collapses under a $50\times$ test-time rescale that a cosine score survives. The rule is one line: score, and compare, reconstruction detectors under a scale-invariant residual; a one-run diagnostic exposes the artifact.

1 INTRODUCTION

Self-supervised vision foundation models such as DINOv2 and DINOv3 (Oquab et al., 2024; Siméoni et al., 2025) promise inspection features without task-specific data, and reconstruction-based unified anomaly detectors built on frozen ViT tokens (You et al., 2022; Guo et al., 2025a;b) are the state of the art on object-centric benchmarks (Bergmann et al., 2021; Zou et al., 2022). On *periodic* manufacturing texture the same features are repeatedly reported to trail purpose-built pipelines (Lei et al., 2024), and one candidate cause suggests itself, and it is the reading we ourselves first adopted: almost every ViT-feature detector ℓ_2-normalizes each patch token before the reconstruction (or distillation, or retrieval) stage, a convention inherited from retrieval (Roth et al., 2022; Deng & Li, 2022), and on wafer, SEM and PCB data removing that single operation moves accuracy by more than any other change. The evidence for that reading is a counterfactual in which one recipe is trained twice with one boolean flipped and nothing else — same frozen backbone, decoder, loss, optimizer, batch size, precision, budget and seed: with normalization on, the wafer detector loses +15.38 mAD (two–three seeds), the SEM detector +38.64, the PCB detector +23.04 (Table 2, left). The natural reading is that normalization erases evidence the head needs.

This paper shows that the reading is wrong, and says where the effect actually lives. The first sign was a control: dividing every token by a *single fixed constant*, an exactly invertible operation that destroys nothing, still cost -12.16 mAD. We then spent a campaign of ablations trying to locate the channel in the training objective — rescaling the decoder's input but not its target (inert), removing the squared-error loss term (the effect shrank by 40% but the baseline fell by twenty points) — and the picture never closed. It closes once one notices what those ablations shared: every one of them was evaluated with the recipe's own anomaly score, the sum of an ℓ_2 residual and a cosine residual between the reconstruction and the (normalized, or rescaled) target. The ℓ_2 term is not scale-

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.

invariant. When the target is unit-normalized, or divided by 49.85, that term no longer measures reconstruction error at a useful scale, and the score, not the model, is what has been degraded. **Rescoring the identical trained checkpoints with the cosine residual alone — no retraining, one configuration line — removes the effect**: on wafer the normalized arm goes from 49.39 to 66.77 mAD, level with the un-normalized arm at 65.87; the fixed-constant arm goes from 54.03 to 65.83; and the two no-squared-error arms, which had read 46.72 and 38.45, both read 65.7–65.8 (Section 4.1). The same rescoring on every dataset of the study is in Table 2.

Three further measurements make this a mechanism rather than a coincidence. First, the decomposition is now clean: with the score held to the cosine residual, the training-side contribution of token scale — loss term present or absent, tokens rescaled or not — is within a point on every arm we trained (Section 4.2). Second, two independent public detectors whose scores are cosine by construction, Dinomaly (Guo et al., 2025a) and INP-Former (Luo et al., 2025), are immune to the identical rescale in both directions, while INP-Former scored with an ℓ_2 residual instead collapses from 70.70 to 52.65 mAD at its native resolution and by 25.4 ± 9.0 over three seeds at 224 px, and the published UniAD checkpoint, whose own score is ℓ_2, loses 22 image-AUROC points under a $50\times$ test-time rescale that leaves a cosine score on the same forward pass unchanged — the same artifact on codebases we did not write (Section 4.5). Third, an attempt to reproduce a training-side effect on INP-Former by adding a squared-error loss term gave a single-seed gain that reversed sign under replication and is retracted; at the weights this recipe uses, the training objective is not where the scale enters.

Contributions.

- A strictly paired counterfactual on six datasets showing the apparent normalization effect under the recipe's own score, and its disappearance under a cosine score on the same checkpoints (Table 2): the effect is an anomaly-score artifact, not information loss.
- The mechanism: an ℓ_2 residual in the score is not scale-invariant, and every control that seemed to implicate the objective — fixed-constant rescale, input-only rescale, no-squared-error loss — is consistently explained once loss and score are separated (Section 4.2).
- A cross-method boundary on three public detectors: two immune when scored with cosine, a 25 ± 9-point collapse over three seeds when one of them is scored with ℓ_2 (only the score changed), the published UniAD checkpoint collapsing under a test-time rescale that a cosine score survives, and a retracted single-seed "loss-side" reproduction, reported with the lesson that seven metrics from one set of predictions are one observation.
- A one-run diagnostic and a reporting rule for the field: rescale the tokens by a constant, or swap the score's residual, and check whether accuracy moves; compare reconstruction detectors under a scale-invariant score.

2 Related Work

Unified anomaly detection on frozen features. UniAD (You et al., 2022) introduced the one-model-per-line setting with an ℓ_2 feature-reconstruction loss and an ℓ_2 anomaly score; HVQ-Trans (Lu et al., 2023), INP-Former (Luo et al., 2025), UniMMAD (Zhao et al., 2026) and the Dinomaly family (Guo et al., 2025a;b) refine the bottleneck and the loss on a frozen DINOv2 backbone and score with cosine residuals. Which residual a method scores with is rarely discussed; we show it decides whether token normalization appears to matter.

Token norms in vision transformers. Large ViTs develop high-norm "register" tokens (Darcet et al., 2024); DINOv3 token norms grow by an order of magnitude across depth (Siméoni et al., 2025). These observations concern the features. Our result is that a scale-sensitive score term can turn any manipulation of those norms into an apparent accuracy effect.

Retrieval-based detection. PatchCore (Roth et al., 2022) retrieves CNN features by Euclidean distance; AnomalyDINO (Damm et al., 2025) retrieves ViT tokens by cosine distance, and few-shot work on DINOv3 finds the metric choice to be at most a development-set effect on object-centric data. Section 4.3 finds the same on our tokens: retrieval reacts little to the metric; the trained head's *score* is where the choice bites.

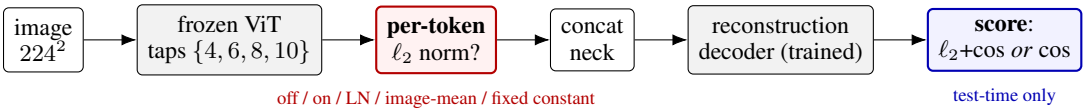

Figure 1: The counterfactual and the score. Every arm of a pair shares the frozen backbone, the trained decoder, the loss, optimizer, batch size, precision, budget and seed; only the red box changes. *off*: raw pre-LayerNorm tokens. *on*: each token divided by its ℓ_2 norm. The blue box is applied to trained checkpoints at test time: the recipe's score is $\|\hat{t} - t\|_2 + (1 - \cos(\hat{t}, t))$ per token; the cosine score drops the first term.

Table 1: Datasets. "Periodicity" is the mean axis-profile autocorrelation strength of the test images (0–1). Training uses normal images only.

Domain	Dataset	Modality / content	Train / test	Budget	Periodicity
Semiconductor / PV	Texture-AD wafer (Lei et al., 2024)	optical AOI, 14 products	10,525 / 4,334	200 ep (65.8k)	0.81
	MIIC circuit SEM (Huang et al., 2021)	SEM, 1 circuit pattern	23,888 / 1,388	7 ep (10.5k)	0.39
	PV cells, EL (Pratt et al., 2023)	electroluminescence, PV cells	144 / 142	1200 ep (10.8k)	0.45
Industrial objects	MVTec-AD (Bergmann et al., 2021)	15 objects/textures	3,629 / 1,725	200 ep (22.7k)	0.72
	VisA-PCB (4 classes) (Zou et al., 2022)	4 PCB products	3,614 / 802	200 ep (22.6k)	0.55
Medical	Medical (4 sub-sets) (Huang et al., 2024)	MRI / US / CT / OCT	3,100 / 6,933	200 ep (19.4k)	—

Periodic-texture inspection. Classical machine vision removes the periodic background in the frequency domain (Tsai & Hsieh, 1999; Aiger & Talbot, 2010; Hou & Zhang, 2007); we make no such claim, and show that the periodic background is not where the normalization effect came from.

3 Setup: One Recipe, One Flag, Two Scores

Pipeline. Figure 1 shows the chain. A frozen ViT (DINOv3-S/16) is tapped at four intermediate blocks; the pre-LayerNorm patch tokens are optionally normalized, concatenated by a 1×1 neck, and reconstructed by a four-layer transformer decoder with a noisy bottleneck and neighbour-masked attention (You et al., 2022). The decoder is trained on normal images only with the loss $0.1\,\|\hat{t} - t\|_2^2 + 0.5\,(1 - \cos)+$ hard-mined cosine, where the tokens t are both the decoder's input and its reconstruction target. This is the published UniAD-style baseline recipe with a DINOv3 prior; it is deliberately not a state-of-the-art decoder.

The flag. `normalize_tokens` is the only field that differs between the two arms of a pair. Under *on*, the normalized tokens are the input, the loss target *and the score reference*; that last role is the one this paper is about.

The two scores. The per-token anomaly map is the recipe's *hybrid* residual, $\|\hat{t}-t\|_2+(1-\cos(\hat{t}, t))$, upsampled bilinearly; the *cosine* score is $1 - \cos(\hat{t}, t)$ alone. Both are evaluated on the same trained checkpoint; switching between them changes one configuration field and no parameter. Rescoring a checkpoint with its own score reproduces its logged number exactly (46.72 on the arm we checked), which is the control for every rescored cell below.

Datasets and protocol. Table 1 lists the six datasets (semiconductor and photovoltaic inspection; industrial objects; a four-modality medical set trained as one unified set following Huang et al. (2024)); all are unified, one model per dataset. Each arm trains for a fixed step budget (Appendix A) and is evaluated at its final epoch; no checkpoint is selected. We report mAD, the mean of seven image- and pixel-level metrics (image AUROC/AP/F1, pixel AUROC/AP/F1, AUPRO), at 224^2 without post-hoc smoothing, so both arms of a pair go through an identical scorer. $\Delta = \mathrm{mAD}_{\mathrm{off}} - \mathrm{mAD}_{\mathrm{on}}$; positive Δ means normalization hurt. Seeds (133, 233, 333) control initialization, data order and bottleneck noise. Seed counts are not uniform across datasets: Texture-AD wafer (optical) three; MIIC circuit SEM three; PV cells (EL) three; MVTec-AD two; VisA PCB1-4 two; Medical (4 sub-datasets) three. Every remaining set is run with one seed. Paired arms always share a seed, so every Δ is a within-seed difference regardless of how many seeds a dataset has.

Table 2: The paired counterfactual under two scores. Same trained checkpoints in both halves; only the test-time residual differs. mAD at the fixed final epoch, mean ± sd over seeds where more than one exists; Δ = off − on, computed within seed. "seeds" gives the number of paired seeds behind each Δ (hybrid / cosine). One recipe, one flag, DINOv3-S/16.

	hybrid score (ℓ_2 + cosine)			cosine score			
Dataset	off	on	Δ	off	on	Δ	seeds
Texture-AD wafer	66.15 ± 0.07	50.77 ± 2.49	**+15.38 ± 2.44**	65.80 ± 0.08	66.88 ± 0.15	**-1.07 ± 0.16**	3 / 3
MIIC circuit SEM	82.34 ± 0.76	43.70 ± 4.54	**+38.64 ± 5.15**	79.88 ± 0.38	79.81 ± 0.06	**+0.07 ± 0.36**	3 / 3
VisA-PCB (4 classes)	72.79 ± 0.27	49.75 ± 3.06	**+23.04 ± 3.20**	70.57 ± 0.68	71.88 ± 0.80	**-1.31 ± 1.34**	3 / 3
MVTec-AD	81.81 ± 0.20	62.29 ± 0.83	**+19.52 ± 0.95**	80.00 ± 0.29	80.29 ± 0.74	**-0.29 ± 0.72**	3 / 3
PV cells, EL	59.92 ± 0.43	60.91 ± 0.49	**-0.99 ± 0.39**	60.58 ± 0.04	61.41 ± 0.49	**-0.83 ± 0.46**	3 / 3
Medical (4 sub-sets)	68.33 ± 0.09	56.24 ± 1.39	**+12.09 ± 1.30**	67.36 ± 0.35	68.83 ± 0.45	**-1.47 ± 0.64**	3 / 3

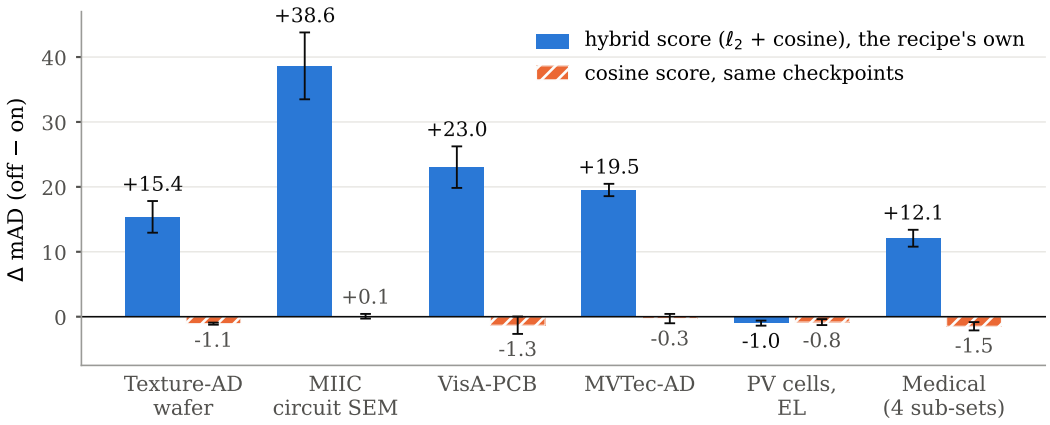

Figure 2: The apparent effect and its disappearance. $\Delta = \mathrm{mAD}_{\mathrm{off}} - \mathrm{mAD}_{\mathrm{on}}$ on the six datasets of Table 2 under the recipe's own hybrid score (solid) and under the cosine residual applied to the *same* trained checkpoints (hatched); mean ± sd over three paired seeds, no retraining. Under the hybrid score normalization costs 12–39 mAD on five of six datasets; under a scale-invariant score every $|\Delta|$ is below 1.5.

4 Results

4.1 The effect, and its disappearance under a cosine score

The left half of Table 2 (solid bars in Figure 2) is the effect as it first appeared: under the recipe's own hybrid score, per-token normalization costs +15.38 ± 2.44 mAD on wafer, +38.64 ± 5.15 on SEM, +23.04 ± 3.20 on VisA-PCB and +19.52 ± 0.95 on MVTec-AD, and nothing on EL (-0.99 ± 0.39). The right half (hatched bars) is the same checkpoints scored with the cosine residual alone. On wafer the normalized arm rises from 49.39 to 66.77 while the un-normalized arm barely moves (66.18 to 65.87), and the paired effect becomes -1.07 ± 0.16; on SEM it becomes +0.07 ± 0.36, on VisA-PCB -1.31 ± 1.34, on MVTec-AD -0.29 ± 0.72, on EL -0.83 ± 0.46 and on the medical set -1.47 ± 0.64. What the cosine score costs the un-normalized arm is -0.31 mAD on wafer: the hybrid score was not extracting anything from the magnitude that the cosine score cannot; it was only losing it when the target's scale changed.

The reason is in the score's definition, and it is not that a larger target makes a larger residual: every metric we report is a ranking metric, so a uniform rescaling of the score map is invisible to all of them. What changes is the map's *shape*. The two residuals are summed without weights: the cosine residual is bounded in $[0, 2]$ and exactly scale-invariant, while the ℓ_2 residual is unbounded and, at the raw token scale (mean token norm 49.85 on wafer), an order of magnitude larger — so at the original scale the hybrid map is effectively the ℓ_2 map, and once the target is unit-normalized

Table 3: Wafer, seed 133: the same checkpoints under three scores. Rows change the tokens or the loss at training time; columns change only the test-time residual. "no-MSE" arms are trained with the squared-error loss term removed. "input-only" rescales the decoder input but not the target or the score reference.

trained arm	tokens	hybrid score	cosine score	ℓ_2 score
off	raw	66.18	65.87	66.21
on	per-token ℓ_2-normalized	49.39	66.77	48.61
fixed constant	$t/49.85$ (input and target)	54.03	65.83	52.83
input-only $\times 1/49.85$	decoder input only	66.45	65.84	—
input-only $\times 50$	decoder input only	66.43	65.87	—
off, rescaled at test time $\times 1/49.85$	input and target, no retraining	38.98	65.19	—
off, rescaled at test time $\times 50$	input and target, no retraining	39.05	65.87	—
off, no-MSE loss	raw	46.72	65.71	46.65
fixed constant, no-MSE loss	$t/49.85$	38.45	65.80	38.46
input-only $\times 1/49.85$, no-MSE loss	decoder input only	45.73	65.95	—

or divided by 49.85 the two terms become comparable and the map is a different mixture. The ℓ_2 residual $\|\hat{t} - t\|_2$ also compares a target at one scale with a network output at its own scale: once the target is rescaled, the term is dominated by the reconstruction's magnitude, which carries no per-token defect evidence, and it no longer ranks pixels as it did. Two channels against one also predicts the sizes we observe: pipelines with an ℓ_2 term in the score move by 15–40 mAD under the flag; a pipeline with such a term only in its loss moves by about 3 (Section 4.5); a pipeline with no such term anywhere should move only through the direction channel, which is small but need not be zero — a prediction Section 4.5 tests rather than assumes.

4.2 Separating loss from score: the controls re-read

Before the rescoring, three controls had been run to locate the channel, and each was read through the hybrid score (Table 3, third column). *Fixed constant*: dividing every token by 49.85, exactly invertible, cost -12.16 mAD — proof that "erased evidence" was not the account, but not yet of what was. *Input-only rescale*: scaling the decoder's input by the same factor while leaving the target untouched cost +0.27 and +0.25, indistinguishable from zero, so the forward pass was not the channel; we concluded the objective was. *No-MSE loss*: removing the squared-error term from the loss shrank the fixed-constant effect from -12.16 to -8.27, "about 40%", but the un-normalized baseline itself fell from 66.18 to 46.72, which we could not explain.

Read through the cosine score (fourth column) the three controls say one thing. The fixed-constant arm is at 65.83, -0.04 from the off arm: the constant costs nothing once the score is scale-invariant. The two no-MSE arms are at 65.71 and 65.80: the twenty-point "collapse" of the no-MSE baseline was the hybrid score failing on a decoder whose output magnitude the cosine-only loss never constrained, and the training-side contribution of removing the squared-error term is -0.16 mAD. The no-MSE off arm replicates across three seeds (46.72, 44.68, 45.86 under the hybrid score; 65.71, 65.66, 65.84 under cosine, a spread of 0.09), so the twenty-point figure and its disappearance are not properties of one run. The other rows replicate on a second seed as well: at seed 233 the fixed-constant arm reads 52.09 under the hybrid score and 65.52 under cosine against 66.21 and 65.83 for the off arm (-14.12 and -0.31), the input-only arm 66.12 and 65.89 (-0.09 and $+0.06$), and the no-MSE fixed-constant arm 38.32 and 66.18 against 44.68 and 65.66 for its no-MSE baseline. The input-only arms were inert under both scores because they never touched the score reference, which is why they looked like evidence for the objective: they had isolated the one path that did not matter. What the squared-error term in the *loss* contributes to the scale effect, with the score held fixed, is the difference between the "fixed constant" and "off" rows under the cosine column, -0.04 mAD; the ℓ_2 column gives the other extreme, the artifact alone. The last two rows separate training from scoring outright: the trained off checkpoint, with its tokens rescaled only at test time, drops to 38.98 and 39.05 under the hybrid score in the two directions and stays at 65.19 and 65.87 under the cosine score, so the collapse needs no retraining at all — it is produced by the score on an unchanged model, and the direction-only channel that survives under cosine is worth at most 0.7 mAD; the seed-233 and seed-333 off checkpoints repeat both rows to within 0.2 (hybrid 38.97 ± 0.07

and 39.06 ± 0.01, cosine 65.33 ± 0.15 and 65.81 ± 0.08 over three seeds), and the ℓ_2-only column repeats too (off 66.21, 66.18 and 66.10 and normalized 48.61, 52.62 and 48.48 over three seeds; fixed-constant 52.83 and 50.97 on two). A weight-only control on a public detector points the same way: setting the squared-error weight of Dinomaly's hybrid loss to $0.1\,c^2$ without touching the tokens reads 71.38 at $c = 0.02$ (against 68.23 for the token-rescaled cell) and 69.81 for $c = 50$ (against 71.04 for the token-rescaled cell and 71.41 for the cosine-only control), so even where a loss-side effect exists it is not the term's relative weight (Section 4.5).

4.3 Where the evidence lives: features, or the head's score?

If magnitude carried the anomaly evidence, one could read it off the frozen encoder without training. A zero-training probe of the token norm separates defective pixels only weakly (pixel AUROC 0.62 on wafer; Appendix F), and nearest-neighbour retrieval on the same tokens reacts little to the metric: switching from cosine to Euclidean distance changes mAD by +0.96 on wafer and -0.51 on SEM (Table 8, Appendix F). The trained head does not extract more from the magnitude either, as the right half of Table 2 shows; what it did was lose accuracy through a score term that reacts to the target's scale.

4.4 Which normalization, and is it optimization?

Table 7 (Appendix D) lists five conventions on the wafer pair under the hybrid score: per-token LayerNorm keeps almost all of the accuracy (+0.84), a per-image scale loses it (-16.30), class-token re-centering is neutral (-0.09). The pattern that puzzled us — LayerNorm keeps relative energy and survives, image-mean scaling keeps relative energy and does not — is the score again: LayerNorm leaves the target at a fixed, image-independent scale, image-mean scaling rescales the target per image. Under the cosine score the conventions read 66.97 (LayerNorm), 66.88 (image-mean) and 65.83 (re-centering), all within 1.0 mAD of the un-normalized arm's 65.87: the convention that loses accuracy under the hybrid score is neutral once the score no longer contains the term it rescales. A learning-rate sweep of the normalized arm (Appendix H) likewise never recovered the gap on SEM and recovered half of it on wafer, because no learning rate repairs a score: the best-tested pair (4×10^{-4}, hybrid 58.12 vs. 66.13) reads 67.18 vs. 65.67 under the cosine score.

4.5 Which published detectors are exposed

Table 4 sorts the family by where an unnormalized squared-error term sits, and it yields three classes rather than two: no such term anywhere (immune), a term in the loss only, and a term in the score (exposed at the headline scale). MambaAD (He et al., 2024) sits in the middle class: its loss is unweighted-scale ℓ_2 (weight 5.0) and its score is the ADer default cosine, code-verified from its released configuration and trainer. We did not run MambaAD, but its class is measured on two independent architectures. Adding an unnormalized squared-error term to an otherwise scale-invariant loss, scored with cosine throughout, costs -2.93, -2.25 and -3.16 mAD at $c = 0.02$ on our Dinomaly variant at weights 0.1, 1 and 10, and -0.30, -0.44 and -0.42 at $c = 50$ (one seed each, pre-registered); with the cosine term removed altogether (pure squared-error loss, cosine score), 68.05 at $c = 1$ against 64.39 at $c = 0.02$, a difference of -3.66, below the $c = 1$ cell in 14 of 14 classes and at the edge of the pre-registered band (-4 to $+0.5$); on an equivalent INP-Former variant the span across $c \in \{0.02, 1, 50\}$ is 4.03 mAD against a matched 0.11 null under the pure-cosine loss (one seed). The loss-side effect is therefore a few mAD, nearly independent of the term's weight and present with or without a cosine term beside it. Two of three seeds of the weight-1, $c = 0.02$ cell diverged during training (seed 133: loss blow-ups at two iterations, final 46.10; seed 333: a blow-up at iteration 3285 in a run still in progress at submission), the only such events in that pipeline; we print the smooth seed's -2.25 and exclude the divergences, since averaging them in would print a number no run produced. A gradient-norm diagnostic on the stable run shows the two loss terms within a factor of 3–12 of each other throughout training, so comparable per-term gradients do not by themselves produce a collapse.

The immune side of the table describes a method family rather than one implementation: RD4AD and Dinomaly, developed independently, score with literally the same expression, $1 - \cos(\hat{t}, t)$, and in the shared ADer benchmark code `cal_anomaly_map` scores with $1 - \cos$ by default and with the unnormalized ℓ_2 residual under `use_cos=False`, so RD4AD, Dinomaly, INP-Former

Table 4: Where an unnormalized ℓ_2 term sits in each detector decides whether a change of token scale can appear as an accuracy change. The property is "no unnormalized squared-error term anywhere", not the score alone: a term in the score produces the large artifact of Table 2; a term in the loss produces a real effect of 2–4 mAD, nearly independent of the term's weight and present with no cosine term at all. Loss/score read from each method's public code or, for MambaAD, from the paper's implementation details; "measured" marks rows we ran (UniAD: its published MVTec-AD checkpoint, no training).

Method	ℓ_2 in loss?	ℓ_2 in score?	under a token rescale
UniAD (You et al., 2022)	yes	yes	exposed (measured on the published checkpoint: a $\times 50$ test-time feature rescale collapses its own ℓ_2 score, image AUROC $96.7 \rightarrow 74.2$, while a cosine score on the same forward pass stays $95.6 \rightarrow 95.7$)
MambaAD (He et al., 2024)	yes (weight 5.0)	no (cosine; ADer default)	loss channel only; not itself measured, but its class is (Dinomaly rows below, including a pure-ℓ_2 loss)
This paper's recipe	yes ($0.1\,\ell_2^2$)	yes (hybrid)	exposed; score term dominant (measured, Tables 2–3)
RD4AD (Deng & Li, 2022)	no	no	immune by construction
Dinomaly (Guo et al., 2025a), as published	no	no	immune (measured: -0.25, -0.07)
Dinomaly $+\lambda\cdot$MSE loss (our variant)	yes	no	loss channel only, cosine-scored. $\times 0.02$: -2.93, -2.25, -3.16 at $\lambda = 0.1, 1, 10$, and -3.66 with the cosine term removed (pure ℓ_2 loss: 68.05 at $\times 1$, 64.39 at $\times 0.02$); $\times 50$: -0.30, -0.44, -0.42 at $\lambda = 0.1, 1, 10$ (measured, one seed each; two of three seeds of the $\lambda = 1$, $\times 0.02$ cell diverged in training and are excluded, see text)
INP-Former (Luo et al., 2025), as published	no	no	immune (measured: $+0.11$, $+0.06$)
INP-Former $+0.1\cdot$MSE loss, cosine-scored	yes	no	loss channel only: span 4.03 vs. a 0.11 null, one seed (measured)
INP-Former, scored with its commented-out ℓ_2 branch	no	yes	collapses -25.4 ± 9.0 mAD (three seeds: -31.4, -15.1, -29.8; measured)

and MambaAD sit in the immune column by one function's default, and the exposed variant is one keyword away. Dinomaly (Guo et al., 2025a) and INP-Former (Luo et al., 2025), each run and scored end-to-end by its own pipeline, move by -0.25 and -0.07 mAD (Dinomaly) and $+0.11$ and $+0.06$ (INP-Former) under the identical fixed-scalar rescale of the encoder tokens ($\times 0.02$ and $\times 50$, training and test; one seed each, seed sd 0.09 and 2.11), against -12.16 for our hybrid-scored decoder and -0.04 for the same decoder scored with cosine: spans of 0.26 and 0.11 mAD over a $2500\times$ range of the applied scale. Adding a squared-error term to their losses does not create a magnitude effect either: across 8 runs the deltas span -1.01 to +0.60 mAD (Table 5), and a single-seed $+2.14$ on INP-Former at 224 px reversed sign under replication ($+2.14$, -1.78, $+0.26$; mean +0.21 $\pm$ 1.96) and is retracted. On Dinomaly's hybrid loss the damage from rescaling tokens is not monotone in the scale (-2.93, -1.00, -0.30 at $\times 0.02$, $\times 1$, $\times 50$; cosine-only control flat over the same range), and the weight sweep above rules out relative weighting.

This immunity is to a scale applied consistently at training and test time, as `normalize_tokens` is; it is not immunity of a fixed forward pass to an off-distribution input. Rescaling only at test time a Dinomaly checkpoint trained at $c = 1$ costs -26.0 mAD at $c = 0.02$ and -42.0 at $c = 50$ under the cosine score, two orders of magnitude more than the 0.26 span held through training, so what confers the immunity is training-time adaptation, not the cosine score by itself. The same test isolates

Table 5: Cross-method test. Each method is run and scored by its own pipeline at its own resolution ($448 \rightarrow 392$ crop, against our 224), so absolute values are not comparable across rows; only the within-method Δ is. In the Dinomaly variants the added squared-error term shares the decoder output with the cosine term, whose hard-mining hook therefore acts on the combined gradient; the recipe is identical in every Dinomaly row.

Method	Setting	baseline	+magnitude	Δ
Dinomaly (CVPR 2024), DINOv2-B/14, $448 \rightarrow 392$ *crop; intervention: loss only (its score is pure cosine)*				
	wafer, DINOv2-B	71.41	70.40	-1.01
	wafer, DINOv2-S	70.11	69.69	-0.42
	SEM, DINOv2-B	91.68	91.10	-0.58
	MVTec-AD, DINOv2-B	89.95	89.80	-0.15
INP-Former (CVPR 2025), DINOv2-B/14, $448 \rightarrow 392$ *crop; intervention: loss and anomaly score (full* 2×2*)*				
	wafer, cos score	70.70	71.26	+0.56
	wafer, cos+0.1mse score	70.68	71.28	+0.60
	SEM, cos score	91.36	91.36	+0.00
	SEM, cos+0.1mse score	91.35	91.42	+0.07

the score channel on a public codebase: on INP-Former's $+0.1\cdot$MSE-loss checkpoints trained at $c = 1$ (three seeds), inflating tokens $50\times$ at test time moves the cosine score by -0.48 ± 0.74 mAD and collapses the ℓ_2 score by -34.94 ± 1.44 on the same forward pass (69.99 to 69.83 and 69.19 to 34.00 on the first seed), whereas shrinking them $50\times$ degrades both (-15.95 ± 11.80 and -40.32 ± 1.70); the $\times 50$ contrast matches our own decoder's test-time rows in Table 3 (65.87 under cosine against 39.05 under the hybrid score, repeated to within 0.2 on two further seeds). On INP-Former's pure-cosine-loss checkpoints the ℓ_2 score is already at 36–56 before any rescale, because nothing constrained the decoder's output magnitude, as in the no-MSE arms of Section 4.2. The published UniAD checkpoint (You et al., 2022), a CNN pipeline whose loss and score are both ℓ_2, behaves the same way with no training on our side: inflating its encoder features $50\times$ at test time leaves a cosine score essentially unchanged (image/pixel AUROC 95.6/94.7 to 95.7/96.7) and collapses its own ℓ_2 score (96.7/96.8 to 74.2/71.1), while shrinking them $50\times$ degrades both (one checkpoint; the unscaled cells reproduce the published numbers). Where a loss-side effect exists under a consistently-applied transform it is a few mAD; on every architecture tested it is an order of magnitude smaller than the score artifact.

The score-side artifact itself transfers. INP-Former's authors left a commented-out ℓ_2 branch in their scoring code; scoring their cosine-trained decoder with it collapses wafer mAD from 70.70 to 52.65 at 392 px, from 67.85 to 36.49 at 224 px, and SEM from 91.36 to 78.63 — one checkpoint per row, only the score changed — while a decoder trained with an added squared-error term holds at 69.17 and 89.02 under the same ℓ_2 score. That is Table 3's no-MSE row on a public codebase: an ℓ_2 residual is a usable score only when the loss has pinned the reconstruction's magnitude, and it never extracts more than the cosine residual does.

4.6 Consequence

The engineering message is not "do not normalize"; it is "do not score with an ℓ_2 residual on a target whose scale you have changed, and do not compare detectors across different residuals". Under a scale-invariant score, normalizing the tokens of a frozen prior is free on every dataset in Table 2, and the un-normalized recipe loses -0.31 mAD by dropping the ℓ_2 term. A constrained decoder on the same frozen DINOv3 — one weight-shared block unrolled eight rounds, hard-mined loose reconstruction targets, magnitudes preserved — reaches 70.60 ± 0.24 mAD at ViT-S and 73.09 ± 0.11 at ViT-B on the wafer benchmark under the Dinomaly protocol with one unified scorer, against 70.54 ± 0.17 and 71.80 ± 0.12 for retrained Dinomaly (Table 6, Appendix I); it scores with a hybrid residual on un-normalized targets, which this paper shows is the configuration in which that residual is not degraded. It also holds under the rule this paper recommends. Rescoring every seed's checkpoint with the cosine main residual (its q_{99} re-estimated on 560 training normals; group maps, fusion and assembly unchanged; the same rescoring with the hybrid residual reproduces the logged 73.06 exactly on the representative seed) gives 72.81 ± 0.11 at ViT-B and 70.49 ± 0.17 at ViT-S

Table 6: Texture-AD wafer under the Dinomaly protocol (256^2 maps, Gaussian $\sigma = 4$, top-1% image score), every row scored by one unified metric implementation; baselines retrained from official code on identical data. mAD is mean $\pm$ sd over seeds (S rows three, B rows four) under each method's own score and, for the magnitude-preserving decoder, after rescoring the same checkpoints with a cosine main residual; Dinomaly's own score is already cosine. AUPRO is the representative seed under the own score.

Model	trainable	mAD, own score	mAD, cosine score	AUPRO
Dinomaly-S (retrained, 3 seeds)	∼15M	70.54 ± 0.17	70.54 ± 0.17 (= own)	.642
Magnitude-preserving ViT-S (3 seeds)	14.5M	70.60 ± 0.24	70.49 ± 0.17	.663
Dinomaly-B (retrained, 3 seeds)	∼35M	71.80 ± 0.12	71.80 ± 0.12 (= own)	.697
Magnitude-preserving ViT-B (4 seeds)	∼20M	$\mathbf{73.09 \pm 0.11}$	$\mathbf{72.81 \pm 0.11}$	.708

(Table 6, fourth column): within 0.3 of the hybrid numbers, still ahead of Dinomaly-B and at parity with Dinomaly-S, whose score is cosine by construction. Every route we tried that let gradients or domain data touch the encoder was neutral or harmful (Appendix I).

5 Scope and Limitations

The rescoring is test-time only and shows a score artifact, not that cosine scoring is optimal. Table 2 uses three paired seeds, whereas the decomposition and loss-side diagnostics are single-seed analyses; the latter show a smaller residual scale effect through the squared-error loss. Our counterfactual covers reconstruction-based detectors only, with all experiments using one backbone at 224^2. The cosine benchmark results are rescored rather than fully retuned models. Finally, the wafer dataset is a public proxy for unavailable production imagery.

6 Conclusion

A large, replicated effect — that ℓ_2-normalizing frozen ViT tokens breaks reconstruction-based inspection detectors on periodic texture — is an artifact of an ℓ_2 residual in the anomaly score, which is not invariant to the scale of its target. Rescoring the same checkpoints with the cosine residual removes the effect on every dataset we tested and re-reads every control we had run in search of a training-side mechanism. Two public detectors scored with cosine are immune; one of them scored with ℓ_2 reproduces the collapse, and the published UniAD checkpoint, whose score is ℓ_2, loses 22 image-AUROC points under a test-time rescale that a cosine score on the same forward pass does not notice. The recommendation is a reporting rule rather than a recipe: compare reconstruction detectors under a scale-invariant score (in the ADer framework, the `use_cos` argument of `cal_anomaly_map`; report both settings), and before attributing an accuracy change to the features, swap the score's residual — one line, no retraining — and see whether the change survives. Ours did not, and we report it.

Reproducibility statement

Every trained arm is a YAML configuration generated by a script that pins all fields except the normalization flag (Appendix A); every rescored cell is the same checkpoint evaluated with one changed field. The configurations, the training and evaluation code, the rescoring configurations and the per-run logs from which every table is built will be released; all numbers are fixed-final-epoch values collected by one script from the logs, and the cross-method runs were executed by the public codebases' own pipelines. The datasets are public; the EL benchmark split is released with the code.

ETHICS STATEMENT

The work uses public inspection and public, de-identified medical benchmark images; no human-subject data were collected. Its practical use is defect screening ahead of human review; we report operating points and false-alarm rates rather than recommending autonomous disposition.

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.

REFERENCES

Dror Aiger and Hugues Talbot. The phase only transform for unsupervised surface defect detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Comput. Vis.*, 129(4):1038–1059, 2021.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Int. Conf. Learn. Represent. (ICLR)*, 2024.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025a.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. One Dinomaly2 detect them all: A unified framework for full-spectrum unsupervised anomaly detection. *arXiv preprint arXiv:2510.17611*, 2025b.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.06564.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Chang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Ling Huang, Deruo Cheng, Xulei Yang, Tong Lin, Yiqiong Shi, Kaiyi Yang, Bah Hwee Gwee, and Bihan Wen. Joint anomaly detection and inpainting for microscopy images via deep self-supervised learning. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2021.

Tianwu Lei, Bohan Wang, Silin Chen, Shurong Cao, and Na Zou. Texture-AD: An anomaly detection dataset and benchmark for real algorithm development. *arXiv preprint arXiv:2409.06367*, 2024.

Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Lawrence Pratt, Jana Mattheus, and Richard Klein. A benchmark dataset for defect detection and classification in electroluminescence images of PV modules using semantic segmentation. *Systems and Soft Computing*, 5:200048, 2023. Dataset: `https://github.com/TheMakiran/BenchmarkELimages` (`dataset_20221008`); ELCell construction protocol in the supplement.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Du-Ming Tsai and Chien-Yu Hsieh. Automated surface inspection for directional textures. *Image Vis. Comput.*, 18(1):49–62, 1999.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

Yuan Zhao, Yunqiu Pang, Longfei Zhang, Hao Liu, Jing Zuo, Huchuan Lu, and Xu Zhao. UniMMAD: Unified multi-modal and multi-class anomaly detection via MoE-driven feature decompression. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

A Configurations and step budgets

Every arm is generated by `tools/make_iclr_configs*.py`. Common recipe: frozen ViT tapped at blocks $\{4, 6, 8, 10\}$ (pre-LayerNorm outputs), 1×1 concatenation neck, four-layer transformer decoder (width 512, 8 heads, feed-forward 1024, 9×9 neighbour-masked attention, noisy-bottleneck dropout 0.2, two spatial adapters of bottleneck 256), AdamW (lr 2×10^{-4}, weight decay 10^{-4}), gradient clipping 0.1, fp32, batch 32 (16 on the two smallest sets), 224^2 inputs with ImageNet normalization, no augmentation. Loss: $0.1\,\ell_2 + 0.5\,(1 - \cos)+$ hard-mined cosine (hardest 25% of tokens, easy weight 0.1). The magnitude-preserving decoder of Table 6 is a weight-shared block of width 1024 unrolled for eight rounds, adapter bottleneck 512, noisy-bottleneck dropout 0.4, taps $\{2, 4, 6, 8\}$, and four loose reconstruction groups through fixed random orthonormal projections (dim 256, weight 0.5); batch 16.

Dataset	epochs	steps	batch
Texture-AD wafer	200	65.8k	32
MIIC circuit SEM	7	10.5k	16
PV cells, EL	1200	10.8k	16
MVTec-AD	200	22.7k	32
VisA-PCB (4 classes)	200	22.6k	32
Medical (4 sub-sets)	200	19.4k	32

B Per-seed tables under the hybrid score

Per-seed values behind the hybrid-score half of Table 2; mAD at the fixed final epoch.

Dataset	backbone/decoder	seed	off	on	Δ
Texture-AD wafer	DINOv3-S	133	66.18	49.39	+16.79
Texture-AD wafer	DINOv3-S	233	66.21	53.64	+12.56
MIIC circuit SEM	DINOv3-S	133	82.98	43.02	+39.96
MIIC circuit SEM	DINOv3-S	233	82.54	39.54	+43.00
MIIC circuit SEM	DINOv3-S	333	81.50	48.55	+32.95
PV cells, EL	DINOv3-S	133	60.28	60.89	-0.61
MVTec-AD	DINOv3-S	133	81.57	62.82	+18.75
VisA-PCB (4 classes)	DINOv3-S	133	72.75	46.90	+25.85
Medical (4 sub-sets)	DINOv3-S	133	68.23	54.87	+13.36

C MEDICAL SUB-DATASETS

Paired effect per medical sub-dataset (unified training on all three). Negative Δ means erasing magnitude helps.

Sub-dataset	seed	off	on	Δ
BUSI	133	61.30	54.12	+7.18
BrainTumor	133	66.86	55.29	+11.57
LiverTumor	133	66.88	62.38	+4.50
Retina	133	77.86	47.68	+30.18

D FIVE NORMALIZATION CONVENTIONS UNDER BOTH SCORES

Table 7: Five conventions on the wafer pair (seed 133, DINOv3-S) under the recipe's hybrid score and, for the same checkpoints, under the cosine score. Each row changes only the red box of Figure 1; the "vs. off" columns are differences to the norm-off arm under the same score. The two conventions that lose 12–16 mAD under the hybrid score (fixed divisor, image-mean) are the two that rescale the reconstruction target; both are within 1.0 of the off arm once the score is cosine.

			hybrid score		cosine score	
Arm	Operation on each token t	keeps relative magnitude	mAD	vs. off	mAD	vs. off
off	none (raw pre-LN tokens)	yes	66.18	–	65.87	–
global-mean (fixed divisor)	$t/\bar{n}$, $\bar{n}$ a dataset constant	yes (exactly)	54.03	-12.16	65.83	-0.04
image-mean	$t/\mathrm{mean}_{\mathrm{image}}\|t\|$	yes	49.89	-16.30	66.88	$+1.01$
LayerNorm	$(t-\mu_t)/\sigma_t$ across channels	mostly	67.03	$+0.84$	66.97	$+1.10$
re-center	$t - t_{\mathrm{cls}}$	no	66.10	-0.09	65.83	-0.04
ℓ_2 (on)	$t/\|t\|$	no	49.39	-16.79	66.77	$+0.90$

E TRAINING DYNAMICS OF THE TWO ARMS

Validation mAD on the sparse checkpoint grid for both arms of each seed-133 pair (values from the run logs; the last column is the fixed final epoch reported in the hybrid-score half of Table 2).

Dataset	arm	trajectory
Texture-AD wafer	off	65.61 → 65.89 → 66.04 → 66.18
Texture-AD wafer	on	54.57 → 50.02 → 50.29 → 49.39
MIIC circuit SEM	off	73.70 → 77.72 → 79.86 → 80.80 → 81.47 → 82.08 → 82.98
MIIC circuit SEM	on	33.50 → 34.29 → 36.34 → 39.80 → 39.15 → 40.91 → 43.02
PV cells, EL	off	60.25 → 60.08 → 60.51 → 60.28
PV cells, EL	on	57.53 → 59.27 → 60.11 → 60.89
MVTec-AD	off	78.74 → 80.04 → 80.60 → 81.57
MVTec-AD	on	60.60 → 63.15 → 62.45 → 62.82
VisA-PCB (4 classes)	off	69.00 → 72.97 → 72.49 → 72.75
VisA-PCB (4 classes)	on	(checkpoint grid not logged for this arm)
Medical (4 sub-sets)	off	66.25 → 67.03 → 67.53 → 68.23
Medical (4 sub-sets)	on	64.43 → 56.19 → 56.47 → 54.87

F TRAINING-FREE PROBES

Table 8: Training-free probes on the same frozen tokens (DINOv3-S, taps $\{4, 6, 8, 10\}$), per-class banks of 64 normal training images. "norm probe": pixel AUROC of $|\,\|t\| - \mathrm{median}\|t\|\,|$. "kNN ℓ_2 / cos": nearest-neighbour anomaly map with Euclidean or cosine distance, mAD; Δ_{kNN} is their difference. Δ_{trained} is the hybrid-score paired effect.

Domain	Dataset	norm probe (P-AUROC)	kNN ℓ_2	kNN cos	Δ_{kNN}	Δ_{trained}
Semiconductor / PV	Texture-AD wafer	0.619	70.64	69.68	+0.96	+14.68
	MIIC circuit SEM	0.797	79.48	80.00	-0.51	+38.64
	PV cells, EL	0.592	66.46	65.94	+0.52	-0.61
Industrial objects	MVTec-AD	0.737	85.54	83.71	+1.83	+18.75
	VisA-PCB (4 classes)	0.750	71.79	66.43	+5.36	+25.85

The zero-training probe of Section 4.3 uses the four tapped pre-LN blocks. A layer-resolved version on wafer and MVTec-AD (blocks 1–12 of DINOv3-S and DINOv2-S, per-image median reference, no smoothing) finds no block at which the raw norm separates defects well: the best DINOv3-S block on wafer reads pixel AUROC 0.56 against 0.53 after the final LayerNorm, and shallow blocks separate object-centric MVTec-AD better than wafer at every depth (up to 0.73 at block 1). The mechanism therefore lives in the head-plus-magnitude pair, not in the norm alone.

G REJECTED PREDICTORS OF THE (HYBRID-SCORE) EFFECT

Before the rescoring, two label-free image statistics were tested as predictors of where the hybrid-score effect appears; neither was validated, and with the effect now attributed to the score the question is moot. The measurements are kept for completeness; Δ is the three-seed hybrid-score value of Table 2. **Rejected candidates.** Two pre-training, label-free image statistics were tested as predictors of Δ and both are rejected on the cleanest available contrast: SEM ($\Delta = 38.64$) and EL ($\Delta = -0.99$) have nearly identical periodicity (0.387 vs. 0.446) and nearly identical local predictability (0.377 vs. 0.390), yet one effect is forty points and the other is zero. Spearman rank correlation with Δ over the 5 datasets with both a finished pair and a periodicity measurement: periodicity -0.40, token-norm CV 0.00 (local predictability has too few finished pairs for a meaningful rank correlation yet; the table below is updated as more pairs finish). Local predictability R is defined in `tools/local_predictability_probe.py`: the fraction of token-norm variance on normal training images explained by a fixed 3x3 spatial neighbour average, at the four tapped pre-LN layers – the exact invariant the neighbour-masked decoder is built to exploit, so its failure to rank Δ is a stronger negative result than periodicity's.

Dataset	periodicity	local predictability R	token-norm CV	Δ
Texture-AD wafer	0.809	0.355	0.064	+15.38
MIIC circuit SEM	0.387	0.377	0.068	+38.64
PV cells, EL	0.446	0.390	0.079	-0.99
MVTec-AD	0.716	0.593	0.074	+19.52
VisA-PCB (4 classes)	0.555	0.415	0.096	+23.04
Medical (4 sub-sets)	—	0.592	—	+12.09

H LEARNING-RATE SWEEP ON THE NORMALIZED ARM

Table 9: Does a lower learning rate on the normalized-token arm recover the hybrid-score gap to the un-normalized arm? The off arm was also run at the on arm's best rate.
Read the trajectory column directly: a declining or flat curve at every lr rejects the optimization-artifact explanation; a curve that recovers toward the off arm at a lower lr supports it.

Dataset	lr	final mAD (vs. off)	trajectory
sem off (reference): 82.98			
	2e-4 (pinned recipe)	43.02 (-39.96)	33.50 → 34.29 → 36.34 → 39.80 → 39.15 → 40.91 → 43.02
	5e-5	36.92 (-46.06)	35.94 → 35.32 → 36.41 → 35.65 → 36.01 → 36.55 → 36.92
	1e-4	38.29 (-44.69)	32.94 → 35.76 → 34.59 → 35.52 → 35.03 → 36.36 → 38.29
	4e-4	46.08 (-36.90)	35.17 → 38.23 → 38.74 → 42.19 → 43.38 → 46.62 → 46.08
wafer off (reference): 66.18			
	2e-4 (pinned recipe)	49.39 (-16.79)	54.57 → 50.02 → 50.29 → 49.39
	5e-5	56.28 (-9.91)	50.43 → 54.04 → 55.70 → 56.28
	1e-4	53.45 (-12.73)	55.38 → 55.64 → 54.41 → 53.45
	4e-4	58.12 (-8.06)	54.64 → 51.12 → 54.69 → 58.12
	4e-4, *off* arm	66.13 (-0.05)	66.16 → 66.26 → 66.10 → 66.13

I BENCHMARK CONTEXT

The rows of Table 6 are our own runs of the constrained decoder (recipe in Appendix A) and of retrained baselines; all are scored under the Dinomaly evaluation protocol (256^2 maps, Gaussian $\sigma = 4$, top-1% image score) by one unified metric implementation applied to every method's raw maps, with baselines retrained from their official code on identical data (three seeds at ViT-S, four at ViT-B). The point relevant here is that once magnitudes are preserved no adaptation of the encoder was needed or helpful: partial unfreezing of the last two blocks cost 2.2 native mAD, and domain-continued self-supervised pretraining cost 16.1. The same single-flag counterfactual re-run in the constrained-decoder codebase reproduces this paper's effect on that decoder family: wafer $66.04 \rightarrow 54.95$ (-11.1, native protocol), VisA-PCB $72.97 \rightarrow 52.12$ at the selected checkpoint and 47.95 at the fixed 100-epoch budget (-20.9 / -25.0), SEM $84.66 \rightarrow 52.07$ (-32.6, epoch 50), with the same monotonically declining norm-on trajectory.