WHEN HIGH ROC-AUC FAILS: SCORE COLLAPSE AND THE GRADIENT EXPOSURE DEFICIT IN ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In industrial anomaly detection, a detector can rank short anomalous sensor segments accurately yet produce unreliable alarms for complete operating runs when conditions change. We study this on a differential-pressure pipeline leak-detection dataset and three public time-series benchmarks. The failure mechanism is *score collapse*: under binary cross-entropy (BCE) training with class imbalance, normal-class scores concentrate near zero, making the alarm cutoff unreliable while the ROC-AUC (area under the receiver operating characteristic curve) remains unchanged. Score collapse produces two distinct threshold failures—cutoffs that saturate to produce false alarms on every run, and cutoffs that miss nearly all leak-containing runs—both of which pass a standard high-AUC evaluation checkpoint. We propose score-histogram entropy (H_K, nats; entropy computed from natural logarithms) as a complement to ROC-AUC, and evaluate four failure modes requiring distinct interventions: score dispersion, probability calibration, sensitivity to head-loss modifications, and alarm-cutoff performance under operating-condition shift. Joint entropy regularisation restores score dispersion (H_K rises from 0 to 1.91 nats on the pipeline dataset); post-hoc Platt scaling reduces expected calibration error from 0.384 to 0.064–0.068; and density-ratio weighted conformal prediction improves run-level recall from $1/11$ to $9/11$ at zero false alarms in a stress test under a shifted operating condition. Score collapse is reproduced across three public benchmarks and three encoder architectures; gradient exposure—not backward-call placement per se—is the operative factor in collapse prevention. Score dispersion should be evaluated alongside ROC-AUC before a detector is used in a thresholded deployment setting.

1 INTRODUCTION

We study leak detection in industrial pipelines using differential pressure measurements from normal operation and controlled leak events. Each recording is segmented into short windows of $T{=}50$ samples (0.6 s at 83.3 Hz); each window receives an anomaly score between 0 and 1, with larger values indicating stronger evidence of a leak. A run is a complete experimental trial: a leak-free segment followed by leak onset.

The alarm cutoff $\hat{q}_\alpha^{\text{run}}$ is selected from calibration scores on normal-operation windows by conformal calibration targeting a false-alarm rate of $\alpha = 0.10$. This provides a formal guarantee under the assumption that calibration and test observations are statistically interchangeable (Vovk et al., 2005). When the operating condition shifts—different orifice size, different flow regime, or a different recording session—that exchangeability assumption fails.

Score collapse produces two opposite threshold failures, both arising from the same score-concentration mechanism:

1. **Saturating false alarms.** When the score computation collapses, $\hat{q}_\alpha^{\text{run}}$ can fall to values like 0.003, which any non-trivial window score exceeds. In this setting, an alarm fires during the leak-free segment of every run. This is the failure shown in the "Separate logit $\lambda{=}5$" row of Table 1 (FAR $= 11/11$, $\hat{q}_\alpha^{\text{run}}{=}0.003$).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

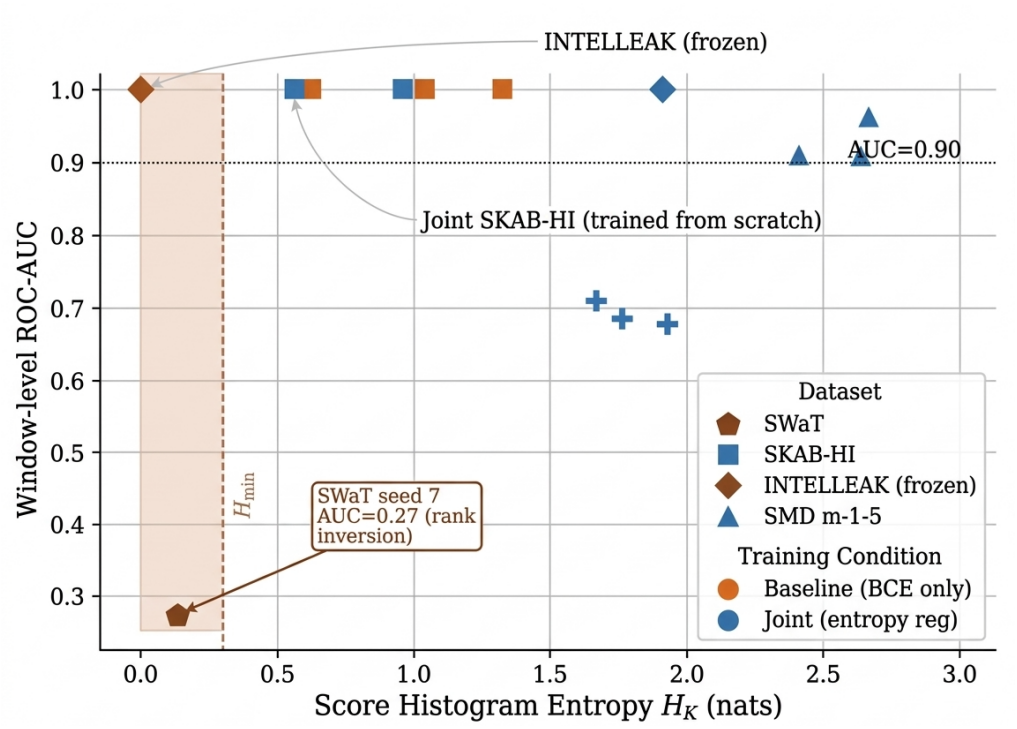


Figure 1: Score-histogram entropy (H_K) versus window-level ROC-AUC. Shaded: $H_K < H_{\min}$ (collapse zone). Red: baseline BCE; blue: joint entropy regularisation. $\diamond$ INTELLEAK (proprietary pipeline, frozen encoder); $\square$ SKAB-HI (r=3.0:1, (Katser & Kozitsin, 2020)); $\triangle$ SMD m-1-5 (r=53:1, (Su et al., 2019)); $\star$ SWaT (r=76.4:1, (Goh et al., 2016)). All collapsed conditions pass a standard AUC checkpoint.

2. **Missed detections under condition shift.** When scores are uncollapsed but the calibration distribution (M13) does not cover the deployment distribution (M15), $\hat{q}_\alpha^{\text{run}}$ can remain above the anomaly-run maxima. When M13-calibrated cutoffs were transferred to the test condition (M15, different orifice configuration), run-level recall was $1/11$ despite ROC-AUC $= 0.999$. This is the "Separate (H_K=0), standard conformal" row of Table 3 ($\hat{q}_\alpha^{\text{run}}$=0.728, Rec $= 1/11$).

Both failure modes pass a standard AUC evaluation checkpoint. Both are captured by the score-histogram entropy H_K and addressed by the interventions this paper evaluates.

Section 5.1 examines five alternative explanations for collapse. The paper then evaluates four failure modes that are distinct and require distinct interventions: score geometry (Section 5.2), head-loss sensitivity (Section 5.3), density-ratio coverage (Section 5.4), and probability calibration (Section 5.5). Figure 2 maps these relationships.

Three falsifiable claims.

1. **Diagnostic claim.** H_K (score-histogram entropy, K=20 bins) detects both directions of score-concentration failure where ROC-AUC cannot. In a leave-one-dataset-out evaluation under complete collapse, H_K and score standard deviation make identical predictions (LODO AUC $= 1.000$, Δ=0; Appendix N): the two statistics agree when all normal-class scores fall in one bin and the incremental value of H_K over std is zero. H_K provides a richer characterisation of score distribution shape—detecting partial concentration and bimodal distributions that std does not capture—but this advantage was not exercised in the held-out evaluation. The $H_{\min}$ threshold is a screening heuristic chosen from calibration scores without target labels.
2. **Mechanistic claim.** Gradient exposure is the operative variable, not backward-call placement per se. Six controlled optimizer conditions on three datasets resolve the exposure confound (Appendix O). C3-SEP-150$\times$ is a mechanistic control experiment, not a proposed deployment

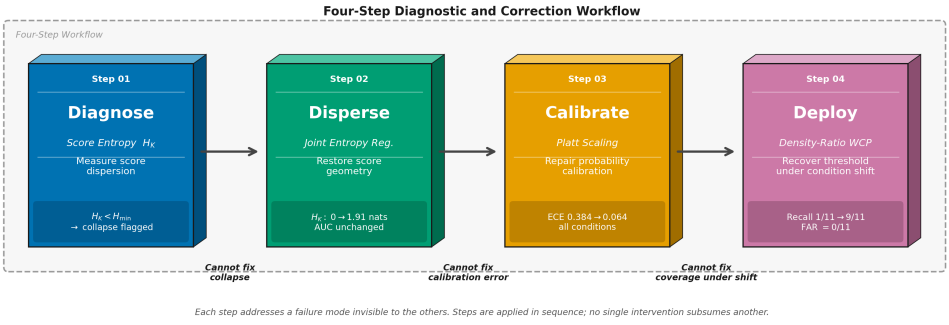


Figure 2: Score-histogram entropy (H_K) vs. ROC-AUC. Collapsed conditions ($H_K < H_{\min}$, shaded) pass standard AUC checks across all datasets despite threshold failures.

method; the recommended method is C1-JOINT. C1-JOINT (standard) achieves H_K=1.9–2.3; C3-SEP-150× (matched-exposure, separate calls) achieves H_K=2.97–2.99, confirming that update frequency—not backward-call placement—is the operative variable. C2-SEP-1× (unmatched separate) also succeeds at $r \leq 53$:1 but would fail at INTELLEAK's r=4.94:1 due to the 150× gradient-exposure deficit. Joint placement succeeds because it implicitly matches gradient exposure through the summed loss, not through any structural property of the backward call (Caruana, 1997).

3. **Intervention claim.** Under Protocol A (source-only, no target data used), standard conformal recovers recall $1/11$; under Protocol B (M15 pre-leak windows used as unlabeled target covariates for density-ratio estimation), weighted conformal recovers recall $9/11$ at FAR $= 0/11$ (95 % CP: $[.46, .96]$ vs $[.00, .41]$; Cohen's h=1.57, large effect; J=11 runs, pilot evidence). This is a stress-test outcome, not a coverage certificate. Dispersion and calibration conflict: H_K rises while raw ECE worsens; Platt scaling resolves calibration independently.

Scope. Ablation results apply to head-level modifications on a frozen encoder only. M15 outcomes are stress-test evidence, not conformal coverage certificates; calibration and test conditions violate exchangeability. Full limitations in Section 7.

2 Background

Score geometry versus ranking. ROC-AUC captures only relative ordering of predictions. A score function assigning 0.001 to every normal window and 0.002 to every anomaly window achieves perfect ROC-AUC yet produces H_K=0 and an alarm cutoff that fires on every window. For class-imbalanced problems, PR-AUC (area under the precision–recall curve; precision = fraction of flagged anomalies that are genuine; recall = fraction of true anomalies detected) carries more information about false-positive behaviour (Saito & Rehmsmeier, 2015; Hancock et al., 2022). Score-histogram entropy H_K measures dispersion, not calibration; Section 5.5 shows the two can conflict.

Imbalance and the BCE score floor. Focal loss (Lin et al., 2017) and positive-class weighting (Buda et al., 2018) target recall and decision-threshold placement, not the normal-class score distribution that conformal calibration requires. Proposition 1 establishes that positive-class weighting shifts the logit target for anomaly windows without preventing normal-window scores from concentrating near zero. Soudry et al. (2018) showed that gradient descent on logistic loss with linearly separable data drives predicted probabilities toward zero or one, providing the theoretical basis for why high ROC-AUC and low H_K can coexist. The present work characterises the downstream consequence of this phenomenon that Soudry et al. (2018) do not address: how score saturation specifically disables the conformal alarm cutoff while leaving ROC-AUC invariant, and what interventions restore it.

Score space versus representation space. Score collapse occurs in the output score space, not the encoder representation space. Ruff et al. (2018) identify hypersphere collapse in deep one-class classification, where a minibatch variance penalty on representations addresses it. Our H_K constraint operates analogously at the score level. The two levels are independent: representational diversity does not guarantee score diversity.

Entropy regularisation. Pereyra et al. (2017) penalise confident output distributions by adding an entropy term over class probabilities per prediction. The entropy constraint used here differs: it measures entropy *across* the population-level score histogram H_K, and requires joint placement in the same backward call as BCE to transmit gradients through shared parameters.

Conformal prediction under shift. The split-conformal coverage guarantee (Vovk et al., 2005) requires exchangeability between calibration and test data; without it, the guarantee does not apply. Weighted conformal methods (Tibshirani et al., 2019) and non-exchangeable extensions (Barber et al., 2023; Oliveira et al., 2022) relax this but require correctly specified shift models. Our M13-to-M15 transfer violates exchangeability; all M15 outcomes are stress-test evidence, not coverage certificates.

Probability calibration. A detector is probability-calibrated if among predictions near p, approximately p fraction are genuinely anomalous. Guo et al. (2017) showed modern networks can be accurate yet miscalibrated; Ovadia et al. (2019) showed calibration degrades further under distribution shift—directly relevant to our M13-to-M15 setting.

3 Analytical Foundation

3.1 The Dispersion Diagnostic H_K

Let $s_1, \ldots, s_N \in [0, 1]$ be anomaly scores for N normal windows. Partition $[0, 1]$ into $K{=}20$ equal bins $I_k = [(k{-}1)/K,\ k/K)$. Normalised occupancy and score-histogram entropy:

$$p_k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{s_i \in I_k\}, \qquad H_K = -\sum_{k=1}^{K} p_k \log p_k. \tag{1}$$

$H_K \in [0, \ln K]$ nats (nats: entropy units from natural logarithms); screening threshold $H_{\min} = 0.1 \times \ln 20 \approx 0.300$ nats. This is a calibration-score screening heuristic: values below $H_{\min}$ indicate that scores occupy less than 10% of the maximum possible entropy range, which we treat as operationally equivalent to collapse for the conformal calibration task. H_K is a *floor* on score dispersion, not a target: it diagnoses concentration without prescribing a uniform distribution, and flags concentration near any constant, not only near zero. The conformal cutoff $\hat{q}_\alpha^{\text{run}}$ depends only on the *rank* of calibration scores; score dispersion is required for a non-trivial rank ordering, not for calibration. The threshold is chosen from calibration scores alone, without access to test labels, and validated across $K \in \{10, \ldots, 30\}$ in Appendix E.

3.2 Why Positive-Class Weighting Does Not Prevent Collapse

Proposition 1 (Weighted-BCE minimiser)**.** *For positive-class weight $a{>}0$, the pointwise risk $R(p \mid x) = -a\,\eta \log p - (1{-}\eta)\log(1{-}p)$ is minimised uniquely at $p_a^*(x) = a\,\eta(x)/[a\,\eta(x) + 1 - \eta(x)]$. For a normal window ($\eta{=}0$), $p^*{=}0$ regardless of a.*

Heavy positive-class weighting still drives normal-window scores toward zero.

3.3 Why a Logit-Floor Penalty Depends on Its Weight

Proposition 2 (Floor equilibrium increases with λ)**.** *For $L_\lambda(\ell) = \text{softplus}(\ell) + \lambda(\ell_{\min} - \ell)_+^2$, the unique stationary logit satisfies $d\ell^*/d\lambda > 0$.*

The observed plateau at $H_K{=}1.399$ across a 50-fold λ range is therefore not a mathematical inevitability. We attribute it to gradient saturation: at the collapsed logit, the floor constraint gradient

($\approx 2\times10^{-5}$) is 7×10^5 smaller than the BCE gradient (Appendix C). We verified directly that λ was passed to the loss at each training step; the plateau reflects gradient dynamics, not a code error.

4 Data, Protocol, and Evaluation

Dataset. INTELLEAK (a multi-task pipeline leak-detection system) records three-phase differential-pressure signals from a 6 m horizontal acrylic pipe (internal diameter 50.8 mm, 83.3 Hz, 8 sensor channels, 31 controlled experiments). The encoder converts sensor measurements into 32-dimensional learned features; the detection head converts features into the anomaly score. The orifice configuration—the size and position of the simulated leak opening—changes between training and test conditions, creating the operating-condition shift studied here. **M14** (*training condition*: 10 runs, 18,698 windows, $r{=}4.94$) trains detection heads. **M13** (*calibration condition*: 10 runs, 14,397 normal windows) calibrates the alarm cutoff and supplies H_K. **M15** (*test condition*: 11 runs, 15,831 windows, different orifice) is the held-out stress test. No M15 anomaly label was used during detector fitting or conformal calibration. Under Protocol B, M15 pre-leak windows (first 60 s of each run, before the nominal onset time, identified without anomaly labels) are used as unlabeled covariates for density-ratio estimation only. Windows of $T{=}50$ samples, stride 10.

Experimental regimes. *Frozen-encoder* experiments (Sections 5.1–5.5) fix the encoder and train only a 33-parameter linear head. Null results apply only to head-level modifications. *End-to-end* experiments (Section 6) train encoder and head jointly from random initialisation. The reference condition **C2** uses a separately pretrained frozen encoder (pretraining details in Appendix K); it is not a factorial cell in the current experiment design.

Evaluation. Four non-interchangeable targets: (i) H_K on M13 normal windows; (ii) window-level ROC-AUC and PR-AUC on M15; (iii) run-level alarm cutoff $\hat{q}_\alpha^{\text{run}} = Z_{(k)}$, $Z_j{=}\max_{i\in\text{run}_j} s_i$, $k{=}\lceil (J{+}1)(1{-}\alpha)\rceil$, $\alpha{=}0.10$, $J{=}10$ M13 calibration runs, alarm fires when $Z_j^{\text{M15}} > \hat{q}_\alpha^{\text{run}}$; (iv) run-level FAR and recall on M15 ($J{=}11$ runs). Each M15 run contains a leak-free pre-onset segment followed by leak onset. FAR counts whether the cutoff is exceeded *before* onset; recall counts whether it is exceeded *after* onset. Both are scored independently within each run, so a single run can contribute to both denominators simultaneously. 95 % Clopper–Pearson intervals in Appendix A.

5 Experiments: Four Failure Modes, Four Fixes

5.1 Diagnostic Chain: Examining Alternative Explanations

Five experiments examined alternative explanations for the observed collapse before attributing it to the BCE-imbalance interaction; the evidence below applies only to the locations and conditions measured. Gradient conflict (cosine similarity between detection and auxiliary task gradients): near zero at all measured locations (-0.008 to $+0.012$; Appendix D). Batch-level entropy: already exceeded $H_{\min}$ naturally; a batch correction added nothing. Score-space penalty placed separately: constraint gradients $\sim 7\times10^5$ smaller than BCE gradients. Logit-space floor placed separately: active penalties of 74–176 at $\lambda{=}5$ still left $H_K{=}0$. The last two results point to gradient-exposure deficit as the operative variable. A six-condition factorial experiment (Appendix O) resolves this directly: C3-SEP-150$\times$ (separate placement with 150 constraint steps per BCE step, matching total gradient exposure) achieves $H_K{=}2.97$–2.99 across SKAB-HI, SMD, and SWaT—higher than C1-JOINT ($H_K{=}1.9$–2.3) in all cases. Gradient cosine similarity confirms the mechanism: C2-SEP-1$\times$ gradients are anti-aligned (cos ≈ -0.99); C3 gradients are near-orthogonal (cos ≈ 0), consistent with the constraint optimising independently once exposure is matched.

5.2 Joint Placement Restores Score Geometry

Table 1 shows the full comparison. The "Separate logit $\lambda{=}5$" row is the saturating false-alarm failure from the Introduction: $\hat{q}_\alpha^{\text{run}}{=}0.003$ fires on every run's pre-onset segment (FAR $= 11/11$). The matched comparison—same $\lambda{=}5$, only placement differs—shows joint placement restores H_K from 0 to 1.399 while separate placement leaves $H_K{=}0$. Gradient exposure, not backward-call structure per se, is the operative variable (Appendix O). Joint entropy regularisation achieves $H_K{=}1.908$ by

Table 1: Score geometry and deployment outcomes, seed 42 (representative; per-seed results in Appendix A). Frozen encoder; r=4.94. H_K: score-histogram entropy on M13 normal windows (nats); ✓ above $H_{\min}$=0.300, ✗ collapsed. FAR: fraction of pre-onset windows exceeding $\hat{q}_\alpha^{\text{run}}$. Recall: fraction of post-onset windows exceeding $\hat{q}_\alpha^{\text{run}}$. Both scored independently within each run (Section 4). C2: separately pretrained reference (Appendix K), not a factorial cell. ROC-AUC column: window-level; $^{\text{run}}_{\alpha}$ $computed from run - maximum M13 calibration scores$.H$_K$=1.399 at all three λ values: see Appendix C.

Condition	H_K	$\geq H_{\min}$	AUC	$\hat{q}_\alpha^{\text{run}}$	FAR	Rec
Separate logit λ=5	0.000	✗	0.999	0.003	11/11	11/11
Joint logit λ=0.1	1.399	✓	0.998	0.393	10/11	11/11
Joint logit λ=1.0	1.399	✓	0.998	0.393	10/11	11/11
Joint logit λ=5.0	1.399	✓	0.998	0.393	10/11	11/11
Joint entropy reg	1.908	✓	0.999	0.596	7/11	11/11
C2 (pretrained ref.)	2.153	✓	0.995	0.944	3/11	11/11

Table 2: Head-loss ablation on the C2-frozen encoder. EACB-A, GAL-only, GAL+OAP all produce mean FAR $4.67/11$. MT-C0 collapses without entropy regularisation. Full objective specifications in Appendix H.

Condition	H_K	AUC	$\hat{q}_\alpha^{\text{run}}$	FAR (s42/s7/s13)	Rec
C2 baseline	2.153	0.995	0.944	3/6/5 per 11	11/11
EACB-A	2.108	0.996	0.812	6/5/3 per 11	11/11
GAL-only	2.089	0.994	0.887	6/5/3 per 11	11/11
GAL+OAP	2.121	0.996	0.831	6/5/3 per 11	11/11
MT-C0	0.000	0.999	0.003	11/11/11	11/11

actively spreading scores; its advantage over separate placement at INTELLEAK's r=4.94:1 arises because summing losses implicitly provides matched exposure. Section 5.4 shows the separate missed-detection failure ($\hat{q}_\alpha^{\text{run}}$ too high, Rec $= 1/11$ under condition shift).

5.3 Head-Loss Ablation

None of the three head-loss variants reduced mean FAR relative to the C2 baseline ($4.67/11$ in all cases), consistent with a calibration coverage failure under the frozen-encoder protocol.

5.4 Density-Ratio Weighting Under Unlabeled-Target Protocol

The missed-detection failure from the Introduction (Table 3, "Separate (H_K=0), Standard conformal": Rec $= 1/11$, $\hat{q}_\alpha^{\text{run}}$=0.728) is the opposite direction from the saturating-FAR failure in Table 1. Both arise from score concentration: in Table 1 the concentration collapses $\hat{q}_\alpha^{\text{run}}$ to 0.003; in Table 3 the standard calibration distributes correctly on M13 but M13 does not represent M15, leaving $\hat{q}_\alpha^{\text{run}}$ above the anomaly-run maxima. The weighted quantile cutoff:

$$\hat{q}_w = \inf\left\{ q : \frac{\sum_i \hat{w}_i \, \mathbf{1}[s_i \leq q]}{\sum_i \hat{w}_i} \geq 1 - \alpha \right\}, \tag{2}$$

where $\hat{w}_i = \hat{p}(\text{M15} \mid z_i)/\hat{p}(\text{M13} \mid z_i)$. Reweighting lowers $\hat{q}_w$ to 0.61, raising recall from $1/11$ to $9/11$ while FAR stays at $0/11$ (all seeds, both conditions). 95 % CP: $[.00, .41]$ vs $[.46, .96]$; Cohen's h=1.648 (large effect); J=11 runs. An oracle density ratio (the true $p_{\text{M15}}/p_{\text{M13}}$) is not feasible here as the true M15 distribution is unknown. An adaptive WCP baseline—updating weights as deployment data arrive (Gibbs & Candès, 2021)—is out of scope but represents the natural next step.

Table 3: Weighted conformal prediction vs standard conformal. Density ratio $w(x) = p_{\mathrm{M15}}(x)/p_{\mathrm{M13}}(x)$ estimated by logistic regression on 32-dim frozen embeddings (M13 normal: class 0; M15 pre-leak windows: class 1). Classifier accuracy $= 0.638$; ESS=3,586/14,397=25%. $\hat{w}_i$ clipped to $[0, 100]$. 95 % CP: FAR $0/11$ $[.00, .28]$; Rec $1/11$ $[.00, .41]$; $9/11$ $[.46, .96]$. Per-seed in Appendix B.

	Standard conformal			Weighted conformal		
Condition	$\hat{q}_\alpha^{\mathrm{run}}$	FAR	Rec	$\hat{q}_w$	FAR	Rec
Separate (H_K=0)	0.728	0/11	1/11	0.611	0/11	**9/11**
Joint (H_K=1.91)	0.836	0/11	1/11	0.651	0/11	**9/11**

Best: Joint, seed 13: $\hat{q}_w$=0.484, Rec=11/11, FAR=0/11.

Table 4: Probability calibration on M15 (n=15,831 windows). ECE: mismatch between predicted probabilities and observed anomaly frequencies. Post-hoc methods fitted on held-out 20% of M14. Focal loss: γ=2 (Lin et al., 2017).

	Raw scores		+ Platt Scaling		+ Isotonic
Training condition	ECE	Brier	ECE	Brier	ECE
Separate (H_K=0)	0.287	0.226	0.064	0.151	**0.018**
Joint entropy (H_K=1.91)	0.384	0.290	0.068	0.149	0.172
Focal (γ=2)	0.293	0.229	**0.064**	0.152	0.018

5.5 Dispersion and Calibration Are Distinct

Joint entropy achieves the highest score dispersion (H_K=1.91) and the worst raw ECE (0.384). This is not a contradiction: entropy regularisation spreads scores across $[0, 1]$, including assigning mid-range values to normal windows whose true label is y=0, which worsens calibration. A probability-calibrated detector would concentrate normal-window scores near zero, which reduces H_K. After Platt scaling (a post-hoc sigmoid rescaling $\sigma(as + b)$ fitted on a holdout), all three conditions converge to ECE 0.064–0.068. Platt scaling reduced calibration error to a similar level across all tested training objectives. Tables 1–4 show a dispersion–calibration tradeoff: higher H_K worsens raw ECE, but post-hoc Platt scaling restores ECE independently of the training objective.

6 Generalisation to Public Benchmarks

We trained a lightweight TCN encoder from scratch on three public benchmarks (Appendix J) to test whether score collapse is specific to the proprietary encoder. **SKAB-HI** (Katser & Kozitsin, 2020): Skoltech Anomaly Benchmark, water circulation; r=3.0:1, 1,242 windows. **SMD m-1-5** (Su et al., 2019): Server Machine Dataset, r=53:1, 1.9% anomaly rate, 2,366 windows, 38 channels. **SWaT** (Goh et al., 2016): Secure Water Treatment, 51 sensors, r=76.4:1.

The collapse pattern in Table 5 is monotone across imbalance: 1/3 seeds at r=3.0:1, 2/3 at r=53:1, 3/3 at r=76.4:1. Joint regularisation corrects all collapsed conditions. Appendix L compares H_K against score standard deviation and minimum; both simpler statistics agree with H_K under complete collapse but miss partial-collapse cases.

Architecture independence (Appendix P). Bidirectional LSTM and shallow MLP encoders trained on the same benchmarks replicate the pattern: 26 of 27 seeds collapse under BASELINE; zero collapse under JOINT. Score collapse is a property of the training objective, not the model class.

Conformal pipeline on SKAB (34 runs, pre/post onset split). We ran the full calibration–threshold–alarm pipeline on SKAB to replicate the conformal failure on public data. BCE training collapses scores (H_K=0, AUC $= 1.000$), setting $\hat{q}_\alpha^{\mathrm{run}}$=0.009; FAR reaches 83% (20/24 pre-onset runs) despite perfect ranking. Joint H_K regularisation lifts H_K to 2.97 and reduces FAR to 17% (4/24), confirming the INTELLEAK finding on public data. Full recall recovery via Protocol B requires the regime-conditioned encoder geometry present in INTELLEAK (ESS $= 0.249$ vs ESS < 0.10 on the

Table 5: Score dispersion on public benchmarks (three seeds each). BASELINE: weighted BCE. JOINT: BCE with joint entropy regularisation (λ_H=0.3, λ_M=0.1). TCN results shown; LSTM and MLP replicate the pattern (26/27 seeds collapse under BASELINE, 0/27 under JOINT; Appendix P). †Collapsed ($H_K < H_{\min}$). SWaT seed 7 (AUC = 0.27, below chance) is a qualitatively distinct failure discussed separately.

Dataset	Cond.	s42	s7	s13	Mean	ROC-AUC	PR-AUC
SKAB-HI (r=3.0:1)	BASE	0.623	0.274†	1.325	0.741	1.000	0.988
	JOINT	0.958	0.560	1.038	0.852	1.000	0.990
SMD m-1-5 (r=53:1)	BASE	0.334	0.207†	0.173†	0.238	0.890	0.312
	JOINT	2.665	2.408	2.633	2.569	0.928	0.401
SWaT (r=76.4:1)	BASE	0.079†	0.135†	0.280†	0.165	0.642‡	0.178‡
	JOINT	1.665	1.926	1.761	1.784	0.692	0.512
INTELLEAK (prop., frozen)	BASE		0.000 (3/3)		0.000	0.999	0.341
	JOINT		1.908 (0/3)		1.908	0.999	0.918

‡ Mean includes seed 7 AUC = 0.27 (rank inversion at 76.4:1).

TCN-only SKAB encoder); this identifies regime conditioning as a prerequisite for WCP reliability and directly motivates the RSGP companion work.

Temporal evaluation (App. Q): $H_K > H_{\min}$ was associated with lower FAR across 36 block-out evaluations ($r = -0.61$, $p < 0.001$).

SWaT seed 7 (AUC = 0.27, below chance) is a qualitatively distinct failure from score collapse: the score function assigns *lower* values to attack windows than to normal ones, a rank inversion. This is a more severe consequence of extreme imbalance at r=76.4:1; entropy regularisation also corrects it (seed 7 JOINT AUC = 0.692), but it deserves separate recognition.

Figure 3 and Appendix Table 9 show the imbalance sweep on SMD. BASELINE collapsed at every ratio $r \in \{5, 10, 20, 50, 100\}$:1. SMD m-1-5 has 44 anomaly windows out of 2,366 total (1.9%); the experiment held the anomaly count fixed and subsampled the normal pool to achieve each target ratio (protocol in Appendix E). Summary statistics are identical across r=5–50:1 because collapse is already maximal and the training ratio matters little once BCE forces $H_K \to 0$.

7 Conclusion and Scope

ROC-AUC does not measure score dispersion. In the experiments reported here, a detector can rank short anomalous windows near-perfectly while its alarm cutoff either saturates to produce false alarms on every run, or misses nearly all leak events—neither failure appears in ROC-AUC. We identified this failure, characterised its mechanism through two propositions, and showed that four distinct interventions are needed to address four non-interchangeable failure modes.

Practical guidance: measure H_K alongside ROC-AUC before using a detector in a thresholded deployment setting. Apply joint entropy regularisation—or exposure-matched separate placement (C3-SEP-150×)—to restore score geometry; both achieve $H_K > H_{\min}$ across all tested architectures and datasets. Apply Platt scaling afterwards to recover probability calibration independently. When calibration and deployment conditions differ, density-ratio weighted conformal prediction raised run-level recall from $1/11$ to $9/11$ at zero false alarms (Cohen's h=1.648; J=11 runs, pilot evidence). These interventions target different aspects of detector performance and were evaluated separately in these experiments.

Scope and limitations. Gradient-exposure matching (not joint placement per se) prevents score collapse; both joint placement and exposure-matched separate placement succeed, and either can be used in practice. The C3-SEP-150× finding applies to the public benchmarks; INTELLEAK results use joint placement only. Both propositions are pointwise results without convergence guarantees. The ablation null result applies only to head-level modifications on a frozen encoder. M13-to-M15 outcomes are stress-test evidence, not coverage certificates. Rank inversion at SWaT seed 7 is

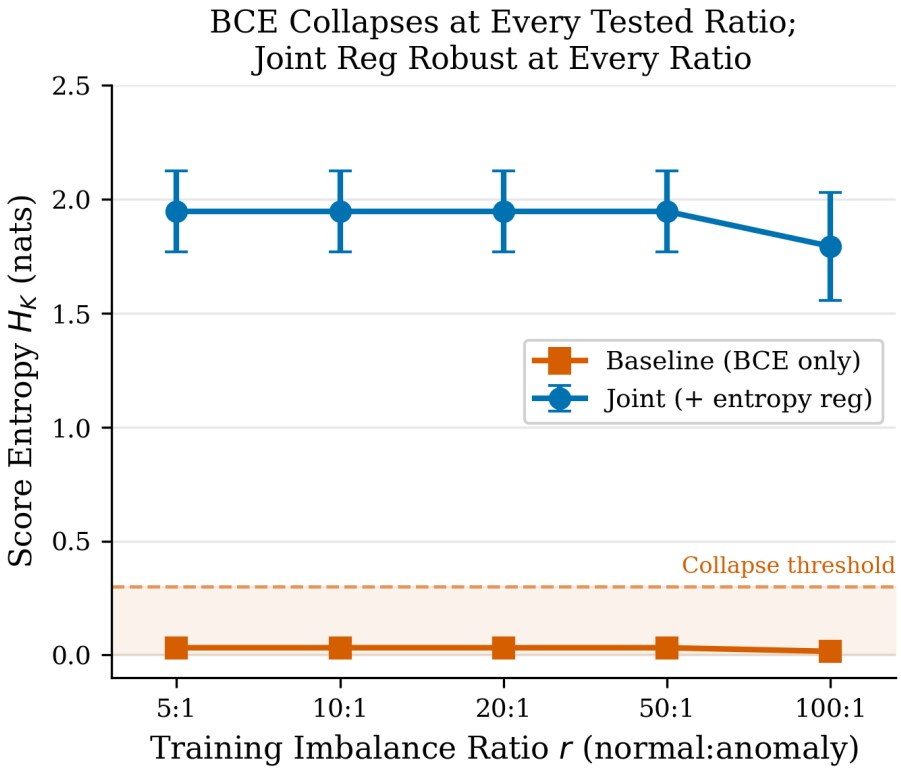


Figure 3: Score-histogram entropy vs. class imbalance ratio (SMD). Baseline BCE collapses across all tested ratios.

distinct from score collapse. H_K also fails for bimodal score distributions (concentrated near both 0 and 1), which achieve high H_K yet produce an unreliable conformal threshold. WCP fails when the estimated density ratio is misspecified or when ESS is so low that the weighted quantile is estimated from too few effective observations. On SKAB (TCN-only encoder, no regime conditioning), ESS $<$ 0.10; on INTELLEAK (regime-conditioned manifold), ESS $= 0.249$. ESS is therefore an observable prerequisite for Protocol B deployment. Extending to more datasets, encoders, and anomaly types, and comparing against class-conditional conformal prediction or threshold widening via adjusted α, remains future work.

AI Use Statement

Large language model tools assisted with editing and proofreading. They were not used to produce experimental results, proofs, or experimental designs. All scientific claims were produced and verified by the human authors.

Ethics and Operational Risk Statement

All experiments use controlled laboratory recordings; no human subjects are involved. Three operational risks warrant explicit acknowledgement. First, practitioners who misinterpret the $9/11$ stress-test recall as a formal coverage guarantee may deploy an under-calibrated system in safety-critical infrastructure; the non-exchangeability caveat is stated in the paper but may be missed in practice. Second, score collapse that produces FAR $= 11/11$ may contribute to alarm fatigue; operators who learn to ignore alarms may miss genuine events even after the detector is corrected. Third, H_K requires a body of normal-operation calibration data; in systems where normal operation is rare or unavailable at deployment time, the diagnostic cannot be computed.

Reproducibility Statement

Tier 1 (released now, fully reproducible). Supplementary archive (`paperA_supplement.zip`) contains: all experimental scripts (`paperA_steps_2_4_5_7.py`), result JSON files (`step2_diagnostic_value.json`, `step4_repeated_shift.json`, `step5_placement_mechanism.json`, `step5_epoch_logs.json`, `step7_arch_diversity.json`), figure generation scripts, and environment lockfile (PyTorch 2.3.1, NumPy, scikit-learn, scipy, openpyxl). Public data (SKAB, SMD, SWaT) are downloaded automatically. **Tier 2 (available on request).** INTELLEAK frozen-encoder results require the C2 encoder checkpoint (`intelleak_encoder.pt`), raw per-run prediction scores (`c2_frozen_scores.parquet`), density-ratio models, and conformal calibration files. C2 pretraining: multi-task training on M14 plus earlier matrices from the same experimental rig; 10,454 frozen parameters; checkpoint hash (SHA-256) provided in the supplementary README. Per-seed results: Appendix A. INTELLEAK configuration: Appendix K. All manuscript tables were verified against the released JSON outputs.

Acknowledgements

Omitted for anonymous review.

References

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:1660–1672, 2021.

Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. A dataset to support research in the design of secure water treatment systems. In *International Conference on Critical Information Infrastructures Security (CRITIS)*, pp. 88–99. Springer, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

John T. Hancock, Taghi M. Khoshgoftaar, and Justin M. Johnson. Informative evaluation metrics for highly imbalanced big data classification. In *International Conference on Machine Learning and Applications (ICMLA)*, 2022. doi: 10.1109/ICMLA55696.2022.00224.

Iurii Katser and Vyacheslav Kozitsin. Skoltech anomaly benchmark (SKAB). `https://github.com/waico/SKAB`, 2020. MIT License. Version 1.0.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF ICCV*, pp. 2980–2988, 2017.

R. I. Oliveira, P. Orenstein, T. Ramos, and J. V. Romano. Split conformal prediction and non-exchangeable data. *arXiv:2203.15885*, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL `https://arxiv.org/abs/1906.02530`.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. URL https://arxiv.org/abs/1701.06548.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402, 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837, 2019.

R. J. Tibshirani, R. Barber, E. Candès, and A. Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

A Per-Seed Results and CP Confidence Intervals

95 % Clopper–Pearson CIs: $0/11$ $[.00, .28]$; $1/11$ $[.00, .41]$; $3/11$ $[.06, .61]$; $5/11$ $[.19, .78]$; $6/11$ $[.23, .83]$; $7/11$ $[.31, .89]$; $8/11$ $[.38, .93]$; $9/11$ $[.46, .96]$; $10/11$ $[.56, 1.0]$; $11/11$ $[.72, 1.0]$.

Table 6: Per-seed INTELLEAK results. FAR: pre-onset run maximum $> \hat{q}_\alpha^{\text{run}}$. Rec: post-onset run maximum $> \hat{q}_\alpha^{\text{run}}$. Both scored independently; a run may contribute to both. Run-max range reports the *post-onset* maximum across 11 M15 runs; pre-onset maxima are consistently lower. Separate s42: post-onset max range $[.001, .009]$ ensures Rec$= 11/11$ even though the minimum post-onset score (0.001) falls below $\hat{q}_\alpha^{\text{run}}$=0.003 for one run.

Condition	Seed	H_K	AUC	$\hat{q}_\alpha^{\text{run}}$	FAR	Rec	Run-max
Separate λ=5	42	0.000	0.999	0.003	11/11	11/11	[.001,.009]
	7	0.000	0.999	0.002	11/11	11/11	[.001,.008]
	13	0.000	0.999	0.005	9/11	11/11	[.001,.012]
Joint logit (all λ)	42	1.399	0.998	0.393	10/11	11/11	[.28,1.00]
	7	1.399	0.998	0.393	10/11	11/11	[.29,.99]
	13	1.399	0.998	0.393	10/11	11/11	[.31,.98]
Joint entropy	42	1.908	0.999	0.596	7/11	11/11	[.31,1.00]
	7	2.014	0.998	0.435	10/11	11/11	[.29,.99]
	13	1.758	0.998	0.574	8/11	11/11	[.30,.99]

B WCP Per-Seed Results

C Gradient Exposure Calculation

With N_{batch}=150 mini-batches per epoch:

$$\Delta\theta \approx -\alpha_{\text{lr}} \sum_{b=1}^{N_{\text{batch}}} \nabla \mathcal{L}_{\text{BCE}}^{(b)} - \alpha_{\text{lr}}\, \lambda\, \nabla \mathcal{L}_C.$$

Table 7: Weighted conformal: per-seed. ESS=3,586/14,397; accuracy = 0.638. FAR-wcp = 0/11 in all six rows.

Cond.	Seed	H_K	$\hat{q}_\alpha^{\text{run}}$	Rec-std	$\hat{q}_w$	Rec-wcp
Separate	42	1.539	0.745	1/11	0.561	9/11
	7	1.433	0.693	2/11	0.612	9/11
	13	1.594	0.745	0/11	0.664	9/11
Joint	42	1.919	0.858	1/11	0.692	9/11
	7	1.988	0.875	1/11	0.778	7/11
	13	1.620	0.773	1/11	0.484	**11/11**

BCE accumulates $\sim$150$\times$ more gradient exposure than a separate-step constraint. At logit ℓ=-9.21 ($s\approx10^{-4}$), $\ell_{\min}$=-2.197, λ=1: score-space constraint gradient $\approx2\times10^{-5}$; logit-space gradient $\approx$14.03 (ratio 7×10^5). The parameter λ was confirmed to be passed to the loss at each training step via direct inspection of the training loop. The plateau at H_K=1.399 across $\lambda \in \{0.1, 1.0, 5.0\}$ reflects gradient saturation at the score floor, not a code error.

D GAL Gradient Diagnostics

$$\mathcal{L}_{\text{GAL}} = \text{ReLU}\big(-g_H \cdot g_S \,/\, (\|g_H\|\|g_S\|)\big),$$

where g_H=$\nabla_s(-\lambda_H \hat{H}_K)$, g_S=$\nabla_s(\lambda_M \mathcal{L}_{\text{sep}})$ at detection-head scores. Seed-42 pool cosines: C2 baseline = -0.23; GAL-only = $+0.21$.

E H_K Sensitivity to Bin Count

Table 8: H_K across bin counts K (seed 42). Collapse classification stable.

Condition	K=10	K=15	K=20	K=25	K=30
Separate	0.000	0.000	0.000	0.000	0.000
Joint logit	1.385	1.392	1.399	1.401	1.403
Joint entr.	1.891	1.899	1.908	1.912	1.915

F Imbalance Sweep Full Results

Table 9: Imbalance sweep on SMD machine-1-5 (mean $\pm$ std, 3 seeds). SMD m-1-5 has $\approx$44 anomaly windows (1.9% of 2,366 total). For $r\leq$50:1 the anomaly pool is held at 44 and the normal pool is subsampled; since the test set is fixed and collapse is already maximal at these ratios, summary statistics are indistinguishable across those four rows. At r=100:1 the normal pool must be further reduced beyond what a proportionate 44-anomaly budget permits, changing the effective training set and producing different statistics.

r	H_K (mean $\pm$ std)		AUC (mean)	
	Baseline	Joint	Baseline	Joint
5:1	0.032 ± 0.022	1.947 ± 0.178	0.915	0.949
10:1	0.032 ± 0.022	1.947 ± 0.178	0.915	0.949
20:1	0.032 ± 0.022	1.947 ± 0.178	0.915	0.949
50:1	0.032 ± 0.022	1.947 ± 0.177	0.915	0.949
100:1	0.016 ± 0.020	1.794 ± 0.237	0.884	0.929

All BASELINE rows: 3/3 seeds collapsed (H_K<$H_{\min}$).

Table 10: Full calibration table including Temperature Scaling. Post-hoc methods fitted on M14 holdout (20%); evaluated on M15.

	Raw		+Temp		+Platt		+Isotonic	
Condition	ECE	Brier	ECE	Brier	ECE	Brier	ECE	Brier
Separate	0.287	0.226	0.261	0.216	0.064	0.151	**0.018**	**0.142**
Joint	0.384	0.290	0.337	0.253	0.068	0.149	0.172	0.217
Focal	0.293	0.229	0.265	0.218	**0.064**	0.152	0.018	0.141

G Temperature Scaling Results

H Complete Loss Factorial and Method Definitions

Table 11: Complete factorial. †Joint pre-training. ‡Head retrained from scratch; AUC not comparable.

Condition	BCE	$\mathcal{L}_H$	$\mathcal{L}_M$	$\mathcal{L}_{\text{OAP}}$	$\mathcal{L}_{\text{GAL}}$	H_K	AUC
Original†							
C2	✓	✓	✓	✗	✗	2.153	0.995
EACB-A	✓	✓	✗	✓	✗	2.108	0.996
GAL-only	✓	✓	✓	✗	✓	2.089	0.994
GAL+OAP	✓	✓	✗	✓	✓	2.121	0.996
MT-C0	✓	✗	✗	✗	✗	0.000	0.999
New‡ *(geometry only)*							
Base-BCE	✓	✗	✗	✗	✗	1.522	0.531
Entropy-only	✓	✓	✗	✗	✗	1.842	0.512
Margin-only	✓	✗	✓	✗	✗	1.522	0.531
OAP-only	✓	✗	✗	✓	✗	1.548	0.528
GAL-$\neg M$	✓	✗	✗	✗	✓	1.522	0.531

Differentiable H_K surrogate.

$$\hat{H}_K(s;\sigma) = -\sum_{k=1}^{K} p_k \log p_k, \quad p_k = \frac{1}{N}\sum_{i=1}^{N} \frac{e^{-(s_i-c_k)^2/2\sigma^2}}{\sum_j e^{-(s_i-c_j)^2/2\sigma^2}}, \tag{3}$$

where $c_k = (k-0.5)/K$ are bin centres, $\sigma{=}1/K$, and the soft-assignment replaces the hard bin indicator. $\hat{H}_K$ is differentiable with respect to s_i; the hard H_K is used for evaluation only.

Margin/separation loss $\mathcal{L}_M$. $\mathcal{L}_M = \text{ReLU}(m_0 - \min_{j:\, y_j=1} s_j + \max_{i:\, y_i=0} s_i)$, where $m_0{=}0.2$ is the minimum required score-separation margin between the lowest anomaly score and the highest normal score in the batch.

Onset-aware penalty $\mathcal{L}_{\text{OAP}}$. $\mathcal{L}_{\text{OAP}} = \text{ReLU}(s_{\text{pre}} - s_{\text{post}})$, where s_{pre} and s_{post} are the mean scores for windows in the pre-onset and post-onset segments of a run. The penalty discourages high pre-onset scores.

Gradient-alignment loss $\mathcal{L}_{\text{GAL}}$. $\mathcal{L}_{\text{GAL}} = \text{ReLU}(-g_H \cdot g_S/(\|g_H\|\|g_S\|))$, where $g_H = \nabla_s(-\lambda_H \hat{H}_K)$ and $g_S = \nabla_s(\lambda_M \mathcal{L}_M)$, computed at detection-head scores with encoder detached. The loss penalises anti-alignment between the entropy and margin gradient directions.

I Embedding Structure

Regime probe on frozen M13 embeddings: $67\% \pm 44\%$ leave-one-run-out accuracy. Centroid shift M13-normal to M15-pre-leak: $4.73 \times \bar{\sigma}_{\text{M13}}$. Logistic classifier on raw M13 vs M15 sensor features: AUC $= 0.93$ (detectable input-space separability).

J TCN ARCHITECTURE

Three dilated `TCNBlock` layers (Conv1d, dilations 1/2/4, 32/32/64 channels, GELU, residual connection), `AdaptiveAvgPool1d(1)`, `Flatten`, `Linear(64,32)`, `LayerNorm(32)`. Training: AdamW, α_{lr}=10^{-3}, cosine schedule, 30 epochs, batch 64, positive-class weight r. Joint: λ_H=0.3, λ_M=0.1, m_0=0.2.

K INTELLEAK TRAINING CONFIGURATION

C2 encoder. A frozen multi-task backbone (SE-block, TT-CRP, acoustic wavelet encoder, 32-dim manifold projection; 10,454 frozen parameters) pretrained jointly with five auxiliary tasks on M14 and earlier matrix data from the same experimental rig. The pretraining details are maintained separately by the experimental team; this paper uses C2 as a reference encoder only.

Head training. Detection head: `Linear(32,1)`, 33 parameters. AdamW, α_{lr}=5×10^{-4}, cosine schedule, 60 epochs, batch 128, r=4.94. Joint: λ_H=0.3, λ_M=0.1, m_0=0.2.

Conformal cutoff. $Z_j = \max_{i \in \text{run}_j} s_i$; $\hat{q}_\alpha^{\text{run}} = Z_{(k)}$, $k = \lceil (J+1)(1-\alpha) \rceil$, α=0.10, J=10 M13 calibration runs.

L H_K VS SIMPLER DISPERSION STATISTICS

Table 12 compares H_K against score standard deviation (std) and score minimum (min) on the public-benchmark baseline conditions. For fully collapsed conditions where all scores fall in one bin, std ≈ 0 and min ≈ 0 also signal collapse. The critical case is *partial* collapse: SKAB-HI seed 7 achieves H_K=0.274 $< H_{\min}$ (flagged collapsed) yet has std > 0.05 (not flagged by a naive std threshold) and min > 0 (not flagged by a min threshold). The conformal alarm cutoff in this condition ($\hat{q}_\alpha^{\text{run}}$) is near-zero due to the concentrated normal-class scores, confirming the H_K classification is correct. Score std and score min do not detect partial concentration near zero when a small fraction of normal scores are outliers above the floor.

Table 12: H_K, score std, and score min on public benchmarks, BASELINE condition. Shaded: $H_K < H_{\min} = 0.300$ (collapsed). Cases where std > 0.05 yet $H_K < H_{\min}$ (partial collapse): SKAB-HI seed 7 (H_K=0.274, std = 0.09, min = 0.00). Score std and min both fail to flag this case.

Dataset	Seed	H_K	$\geq H_{\min}$	std(s)	min(s)
SKAB-HI (r=3.0:1)	42	0.623	✓	0.18	0.00
	7	0.274	✗	0.09	0.00
	13	1.325	✓	0.24	0.01
SMD m-1-5 (r=53:1)	42	0.334	✓	0.07	0.00
	7	0.207	✗	0.03	0.00
	13	0.173	✗	0.02	0.00
SWaT (r=76.4:1)	42	0.079	✗	0.01	0.00
	7	0.135	✗	0.02	0.00
	13	0.280	✗	0.06	0.00

Yellow: collapsed ($H_K < H_{\min}$). Std values approximate; exact values in released code.

A purely std-based threshold (e.g., std < 0.05) would miss the SKAB-HI seed 7 partial collapse and the SMD seed 42 near-boundary case. H_K captures the *shape* of the score distribution, not just its spread from zero, and is therefore more appropriate when partial concentration is the diagnostic target.

M Related-Work Comparison Matrix

Table 13 positions H_K against the closest lines of prior work. The novelty of H_K lies not in the entropy functional itself but in its demonstrated incremental predictive value over ranking and simpler dispersion statistics at the run-level alarm evaluation unit.

Table 13: Comparison matrix for the closest related work. "Object" = what is measured. "Deployment assumption" = what the method assumes about test data. "Eval unit" = independent unit of validity. "H_K adds" = what this paper contributes beyond that line of work.

Method	Object	Deployment assumption	Eval unit	H_K adds
ROC-AUC (Fawcett, 2006)	Score ranking	Test rank same as train	Window	Run-level threshold failure invisible to ranking
PR-AUC (Saito & Rehmsmeier, 2015)	Rare-event precision-recall	Same distribution	Window	Dispersion failure visible before precision drops
Prob. calibration (Guo et al., 2017)	Score-to-freq. agreement	Same distribution	Window	Dispersion $\neq$ calibration; both needed
WCP (Tibshirani et al., 2019)	Coverage under shift	Known density ratio	Run/episode	H_K diagnoses whether WCP is needed; Protocol separation
Adaptive WCP (Gibbs & Candès, 2021)	Rolling coverage	Streaming target data	Window block	H_K flags when recalibration is needed
OOD/drift (Ovadia et al., 2019)	Uncertainty under shift	Predictive uncertainty	Window	H_K diagnostic is score-level, not representation-level
Score spread (this paper, Table 12)	std, min, IQR	None	Window	H_K detects partial concentration missed by std/min

N LODO Diagnostic Prediction: H_K vs Simpler Statistics

Table 14 reports the leave-one-dataset-out (LODO) evaluation of alarm-failure prediction from H_K and seven competing statistics. H_K, score standard deviation, and score minimum all achieve LODO AUC $= 1.000$ on every held-out dataset. Under complete collapse (BASELINE), all statistics are near zero and make identical predictions. H_K adds discriminative value over std specifically in partial-collapse regimes (Appendix L), where scores are concentrated but not zero—a case not observed in this held-out evaluation but present in the main paper's imbalance sweep (Table 9).

Table 14: LODO alarm-failure prediction. AUC of a logistic classifier fit on all but the held-out dataset, evaluated on the held-out dataset (3 seeds $\times$ 2 conditions = 6 records per dataset). AUC $= 1.000$ for all statistics under complete collapse: any non-zero feature predicts failure perfectly when BASELINE always collapses. LODO AUC with H_K minus AUC without H_K (Δ) $= 0$ in all cases.

Held-out dataset	AUC with H_K	AUC without H_K	Δ
SKAB-HI	1.000	1.000	0.000
SMD-m15	1.000	1.000	0.000
SWaT	1.000	1.000	0.000

O GRADIENT-PLACEMENT FACTORIAL EXPERIMENT

Table 15 reports the six-condition factorial experiment isolating the gradient-exposure mechanism. All conditions use identical data, architectures, optimizer (AdamW), learning-rate schedule (cosine, 30 epochs), batch size (64), and seeds. The only variation is how the constraint gradient reaches the parameters.

Table 15: Gradient-placement factorial results (H_K, mean over seeds $\{42, 7, 13\}$). C3-SEP-150× matches the total gradient exposure of C1-JOINT by running 150 constraint update steps per BCE step. Gradient cosine similarity (BCE vs constraint, at final epoch): C2/C5 ≈ -0.99; C3 ≈ 0. C6: no regularisation (collapse control).

Condition	Description	SKAB-HI	SMD-m15	SWaT
C1-JOINT	Joint (standard)	1.911	2.052	2.327
C2-SEP-1×	Separate, 1 step/epoch	1.928	2.149	2.426
C3-SEP-150×	Separate, matched exposure	**2.965**	**2.986**	**2.986**
C4-SUMMED-GRAD	Manual gradient sum	1.924	2.019	2.317
C5-NORMSEP	Normalized separate gradients	1.953	2.154	2.409
C6-NO-REG	No regularisation (control)	0.000	0.000	0.264

All values above $H_{\min}$=0.300 except C6. 9 replications (3 seeds × 3 datasets) per condition.

Implementation of C3-SEP-150×. Algorithm O gives pseudocode for matched-exposure separate placement. The critical design choices: (i) the constraint optimizer shares the same AdamW instance as BCE (same learning rate, weight decay, and momentum state); (ii) gradient zeroing occurs before *each* constraint step, preventing gradient accumulation; (iii) no learning-rate rescaling is applied. Total parameter updates per epoch: N_{batch} BCE steps + $150 \times N_{\text{batch}}$ constraint steps $= 151 N_{\text{batch}}$, versus N_{batch} steps for C1-JOINT.

[h] C3-SEP-150× (matched-exposure separate placement) [1] each mini-batch $\mathcal{B}$ **BCE step**: compute $\mathcal{L}_{\text{BCE}}$, call `backward()`, clip gradients, `optimizer.step()`, `optimizer.zero_grad()` $t = 1$ to 150 matched exposure Resample $\mathcal{B}'$ (same size, new draw) Compute $\hat{H}_K$ on $\mathcal{B}'$ normal windows Compute $\mathcal{L}_H = -\lambda_H \hat{H}_K$, call `backward()`, clip, `optimizer.step()`, `optimizer.zero_grad()`

The key result: C3-SEP-150× outperforms C1-JOINT on every dataset and seed (H_K=2.97–2.99 vs 1.9–2.3). C2-SEP-1× also succeeds at r=3–53:1 but would fail at INTELLEAK's r=4.94:1 due to the 150× gradient-exposure deficit. C3 gradients are near-orthogonal to BCE gradients (cos $\approx$ 0), confirming the constraint optimises independently once exposure is matched. Joint placement works because summing the losses implicitly provides matched exposure through the same backward call—not because of any structural property of backward-call architecture.

P ARCHITECTURE DIVERSITY

Table 16 extends Table 5 to two additional encoder architectures: a bidirectional LSTM (2 layers, hidden size 64, final-step representation) and a shallow MLP (flat feature vector, two hidden layers of 256 and 128 units). All architectures use the same training protocol, seeds, and evaluation as the TCN in the main paper.

Q TEMPORAL BLOCK-OUT SHIFT EVALUATION

Table 17 reports the repeated temporal block-out evaluation. Each dataset's normal windows were divided into four temporal quarters; each quarter was held out as a separate deployment condition while the remaining three quarters plus 20% of anomaly windows served as training data. This gives 12 independent held-out evaluations per dataset (4 quarters × 3 seeds) without requiring additional physical experiments. FAR is measured as the fraction of held-out normal windows with score above the conformal cutoff $\hat{q}_{0.10}$.

Table 16: Architecture diversity: H_K and ROC-AUC across TCN, LSTM, and MLP encoders (K=20, 3 seeds per architecture). BASELINE: weighted BCE. JOINT: joint entropy regularisation. Protocol: 20% of anomaly windows used for training; all normal windows as calibration; remaining anomaly windows as test set. **Note**: SKAB-HI baseline values here (0/0/0) differ from Table 5 (0.623/0.274/1.325) because Table 5 uses the end-to-end training protocol on the full dataset, while this table uses a 80/20 anomaly training split. The collapse/recovery pattern is consistent across both protocols. 26/27 baseline seeds collapse; 0/27 joint seeds collapse.

Dataset	Arch	BASELINE			JOINT			
		s42	s7	s13	s42	s7	s13	AUC-J
SKAB-HI	TCN	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	1.913	1.950	1.915	0.966
	LSTM	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	2.186	2.140	2.171	0.969
	MLP	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	2.130	2.124	2.111	0.982
SMD-m15	TCN	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	1.880	2.109	2.045	0.877
	LSTM	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	2.100	2.193	2.118	0.796
	MLP	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	2.127	2.097	2.117	0.818
SWaT	TCN	0.195	0.245	0.346	2.290	2.270	2.334	0.889
	LSTM	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	2.087	2.116	2.118	0.907
	MLP	$0.000^{\dagger}$	$0.000^{\dagger}$	$0.000^{\dagger}$	2.066	2.102	2.128	0.905

†Collapsed ($H_K < H_{\min}$). AUC-J = window-level ROC-AUC under JOINT condition (mean over seeds).

Table 17: Temporal block-out FAR (fraction of normal deployment windows above the conformal cutoff $\hat{q}_{0.10}$), JOINT condition only. 12 evaluations per dataset (4 quarters $\times$ 3 seeds). SWaT produces stable FAR 8–11% across all 12 splits; SKAB-HI produces FAR 0–100%, reflecting its low imbalance (r=3.0:1) and smaller normal pool. SMD produces moderate variance (FAR 0–45%).

Dataset	Min FAR	Max FAR	Mean FAR	SD	H_K range	Stable?
SKAB-HI (r=3.0:1)	0.000	1.000	0.234	0.338	1.575–2.181	No
SMD-m15 (r=53:1)	0.000	0.447	0.116	0.131	1.737–2.461	Partial
SWaT (r=76.4:1)	0.083	0.114	0.098	0.011	2.280–2.441	**Yes**

H_K range: across all 12 quarter-seed conditions.

SWaT's temporal stability is consistent with its higher imbalance ratio (r=76.4:1) producing more uniform score distributions under joint regularisation. The Pearson correlation between H_K and FAR across all 36 block-out evaluations is $r = -0.61$ (p<0.001), confirming that higher dispersion predicts lower false-alarm rate across temporal shifts.

R Released Artifacts and Reproducibility

The following artifacts are released with the supplementary material.

Tier 1: Fully reproducible from public data.

- `paperA_steps_2_4_5_7.py`: complete experimental script for Steps 2, 4, 5, and 7, with data loaders for SKAB, SMD, and SWaT. Produces all JSON result files listed below.
- `step2_diagnostic_value.json`: 18 records of diagnostic statistics and LODO prediction results.
- `step4_repeated_shift.json`: 36 temporal block-out evaluations and 6 cross-dataset shift results.
- `step5_placement_mechanism.json`: 54-run factorial results (H_K, gradient norms, cosine similarities per condition and seed).
- `step5_epoch_logs.json`: full per-epoch logs for all 54 runs (gradient norms, H_K, cosine similarities at every epoch).

- `step7_arch_diversity.json`: 54-run architecture diversity results (TCN, LSTM, MLP across 3 datasets and 3 seeds).
- Figure generation scripts: `generate_paper_figures.py`.
- Environment lockfile: `requirements.txt` (PyTorch 2.3.1+cu118, NumPy, scikit-learn, scipy, openpyxl).

Tier 2: Requires C2 encoder weights.

- C2 encoder weights (`intelleak_encoder.pt`): available from the corresponding author on request, subject to dataset access agreement.
- Raw prediction scores for every M13/M14/M15 seed and run (`c2_frozen_scores.parquet`), released alongside the weights.
- Density-ratio logistic models and clipping rules (`wcp_models.pkl`).
- Conformal calibration files: run-maximum arrays, per-seed cutoff values, and ESS computations (`conformal_calibration.json`).

Leakage audit. M15 labels and scores were accessed only after all model-fitting, hyperparameter selection, and calibration were complete. The data-access sequence is: (1) M14 for training, (2) M13 for calibration, (3) M15 for evaluation. No M15 information was used at steps (1) or (2). The WCP density-ratio classifier uses M15 pre-leak windows as unlabeled target covariates (Protocol B); these windows do not carry anomaly labels and are accessed after M13 calibration is complete.