WHEN SHOULD ANOMALY THRESHOLDS ADAPT? A DECISION FRAMEWORK UNDER PREVALENCE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised anomaly detectors produce continuous scores, but deployment requires thresholds that convert those scores into decisions. Reference-calibrated thresholds can become unreliable when anomaly prevalence changes, while indiscriminate batch adaptation can introduce error when the score distribution provides little evidence that adaptation is warranted. We introduce **CAVET-AD** (**C**alibrated **A**daptive **V**alue of **E**xcess **T**ail for **A**nomaly **D**etection), a detector-agnostic post-hoc framework that determines whether threshold adaptation is supported by the observed score distribution. CAVET-AD calibrates scores against a frozen nominal reference, refines the upper tail using extreme-value theory, and computes an observable excess-tail response, r_{tail}, that controls a bounded adaptation budget. When the response is near zero, the method retains its conservative base rule rather than forcing adaptation. Across four tabular benchmarks and 320 positive-prevalence conditions, CAVET-AD increases descriptive mean F1 from 0.475 to 0.579. Its behavior is strongly regime dependent: gains are substantial on CICIoT2023 and CreditCard2013, NSL-KDD shifts from fixed-favorable to adaptive-favorable as prevalence increases, and CICIDS2018 shows no benefit when Isolation Forest scores are nearly uninformative. At the natural CreditCard2013 prevalence of 0.173%, adaptation is minimal and yields no statistically significant F1 improvement after multiplicity correction. A leave-one-dataset-out policy based on r_{tail} retains most of the adaptive benefit on a benchmark excluded from cutoff selection, while ECOD replication and detector-transfer analyses provide additional evidence of portability. Under severe controlled covariate shift, however, r_{tail} can increase even when adaptation is harmful. These results position CAVET-AD as a diagnostic for when prevalence-driven threshold adaptation is warranted, rather than as a universal detector of distribution shift.

1 INTRODUCTION

Unsupervised anomaly detection is widely used in fraud analytics, network intrusion detection, IoT telemetry, and industrial monitoring. A detector assigns each instance a scalar anomaly score $s(x)$, but deployment ultimately requires an operating threshold that converts scores into binary decisions. In practice, this threshold is often calibrated using a target contamination rate, a score quantile, or a validation operating point. Yet deployment anomaly prevalence π may differ substantially from the calibration regime, may change over time, and may even exceed 0.5. Under such prevalence shift, a fixed threshold can become poorly calibrated (Godau et al., 2025; Ruff et al., 2021).

Simply adapting the threshold to every incoming batch is not a satisfactory solution. When the deployment score distribution contains little usable evidence of prevalence change, adaptation can introduce error rather than correct it. This creates a more fundamental deployment question: not merely *how* should an anomaly threshold adapt, but *when* is adaptation actually warranted?

Problem statement. Can an unsupervised post-hoc calibration scheme (i) adapt its operating point when the deployment score distribution provides evidence consistent with prevalence change, and (ii) recognize when such adaptation is not warranted, without requiring labeled deployment data, knowledge of the true deployment prevalence, or detector retraining?

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.

We introduce **CAVET-AD** (Calibrated Adaptive Value of Excess Tail for Anomaly Detection), a detector-agnostic post-hoc framework designed around this question. CAVET-AD maps anomaly scores into a calibrated reference-tail space and measures the label-free excess-tail response r_{tail}, which quantifies additional upper-tail mass relative to a frozen nominal reference. This observable response controls a bounded adaptive operating budget from which the decision threshold is obtained. When r_{tail} is near zero, CAVET-AD retains its conservative base rule rather than forcing adaptation. Thus, the same observable quantity both drives adaptation and provides a deployment-time indication of whether adaptation is supported by the score distribution.

This decision behavior distinguishes CAVET-AD from approaches that always adapt and from methods that require deployment labels or explicit prevalence estimation to determine the operating point. Importantly, r_{tail} is not assumed to estimate the true anomaly prevalence, and CAVET-AD is not intended as a universal diagnostic for arbitrary distribution shift. We explicitly test this boundary under controlled covariate shift.

We evaluate CAVET-AD on four heterogeneous tabular benchmarks using a frozen paired ten-seed confirmatory protocol with Isolation Forest, followed by a full-protocol replication with ECOD using the same datasets, seeds, prevalence grid, partitions, and calibration configuration.

1.1 Contributions

CAVET-AD makes three main contributions:

(1) A decision rule for adaptive operating-point adjustment. Under prevalence shift, the label-free statistic r_{tail} simultaneously drives threshold adaptation when excess tail mass is present and identifies a no-adaptation boundary when it is absent. This provides a deployment-time applicability signal without labeled deployment data, knowledge of the true prevalence, or detector retraining. A controlled covariate-shift stress test identifies an important boundary at which this interpretation ceases to be reliable.

(2) A detector-portable reference-tail calibration framework. CAVET-AD maps oriented anomaly scores through a frozen empirical reference CDF, refines the upper tail using extreme-value theory, and obtains the operating threshold through survivor-function inversion. The construction separates detector scoring from threshold adaptation and can therefore be applied post hoc to continuous-score anomaly detectors without retraining them.

(3) A conditional, replicated, and held-out evaluation. A frozen paired evaluation across four benchmarks and controlled prevalence conditions demonstrates that the benefit of adaptation is regime dependent. A full-protocol ECOD replication and supplementary LOF analysis examine detector portability, while a leave-one-dataset-out policy tests whether r_{tail} can select between fixed and adaptive rules without using the held-out dataset to fit the decision cutoff. The resulting evidence distinguishes regimes in which adaptation helps, becomes preferable only after a prevalence-dependent crossover, or should be withheld.

2 Background and Related Work

Score-based anomaly detectors, including Isolation Forest, LOF, One-Class SVM, ECOD, KDE, mixture models, and deep reconstruction methods, produce continuous scores that require a separate operating rule (Liu et al., 2008; Breunig et al., 2000; Schölkopf et al., 2001; Li et al., 2023; Parzen, 1962; Bishop, 2006; Zhou & Paffenroth, 2017). EVT-based methods model extreme-score tails (Siffer et al., 2017; Li et al., 2021; Davis et al., 2020), but still require an operating risk or exceedance level. CAVET-AD instead uses EVT to refine a frozen reference tail and adapts the operating budget from an observable excess-tail response.

Adaptive conformal methods target coverage under exchangeability or specific shift models (Vovk et al., 2005; Gibbs & Candès, 2024; Martinez Gil et al., 2026), whereas prevalence-estimation methods recover a deployment mixture proportion under additional identifiability assumptions (Scott, 2015; Jain et al., 2016; Perini et al., 2022; 2023). CAVET-AD instead adapts the operating point without estimating prevalence or prescribing a coverage target, following reference calibration $\rightarrow$ $r_{\mathrm{tail}} \rightarrow b(r_{\mathrm{tail}}) \rightarrow \tau^*$. The same r_{tail} supplies the deployment-time applicability diagnostic evaluated in Section 5.4; detailed literature positioning appears in Appendix A.

3 METHODOLOGY: CAVET-AD

CAVET-AD is a detector-agnostic post-hoc decision framework summarized by

$$S_{\mathrm{ref}} \xrightarrow{\widehat{F}_{\mathrm{ref}}} Z_{\mathrm{ref}} \xrightarrow{\mathrm{EVT}} \widehat{\overline{F}}_{\mathrm{ref,EVT}},$$

$$S_{\mathrm{batch}} \xrightarrow{\widehat{F}_{\mathrm{ref}}} Z_{\mathrm{batch}} \xrightarrow{\text{tail excess}} r_{\mathrm{tail}} \longrightarrow \begin{cases} \text{applicability diagnostic,} \\ b(r_{\mathrm{tail}}) \xrightarrow{\widehat{\overline{F}}^{-1}_{\mathrm{ref,EVT}}} \tau^*. \end{cases}$$

The following subsections define each component.

3.1 PROBLEM SETUP AND NOTATION

Let observations follow

$$x_i \sim (1-\pi)P_0 + \pi P_1,$$

where P_0 and P_1 denote the normal and anomalous distributions and π is the unknown anomaly prevalence. A base detector produces oriented scores $s_i = s(x_i)$, with larger values indicating stronger anomaly evidence. CAVET-AD transforms these scores into a reference-tail Z-space, measures the label-free excess-tail response r_{tail}, and derives $\tau^*(r_{\mathrm{tail}})$. The true prevalence π is not an input to the deployed rule and is varied only for controlled evaluation.

3.2 REFERENCE-SCORE-TO-Z TRANSFORMATION AND TAIL CALIBRATION

Because detector scores can differ in scale and shape, CAVET-AD maps oriented scores to a common reference-tail Z-space.

Step 1: Estimate the reference score CDF. Let $\mathcal{S}_{\mathrm{ref}}$ denote the oriented score sample available during calibration and let $N_{\mathrm{ref}} = |\mathcal{S}_{\mathrm{ref}}|$. CAVET-AD uses the finite-sample empirical CDF

$$\widehat{F}_{\mathrm{ref}}(s) = \frac{\#\{s_j \in \mathcal{S}_{\mathrm{ref}} : s_j \le s\}}{N_{\mathrm{ref}} + 1}.$$

The $N_{\mathrm{ref}} + 1$ denominator provides a finite-sample plotting-position adjustment (Coles, 2001) that keeps $\widehat{F}_{\mathrm{ref}}(s) < 1$, ensuring the log-transformation $Z = -\log(1 - \widehat{F}_{\mathrm{ref}}(s))$ remains finite. Scores exceeding all reference values map to $Z = \log(N_{\mathrm{ref}} + 1)$ and enter the EVT-refined tail region.

Step 2: Transform to reference-tail Z-space.

Each score maps to

$$Z_i = -\log\left(1 - \widehat{F}_{\mathrm{ref}}(s_i)\right).$$

Under adequate reference calibration, nominal Z values should approximately follow an $\mathrm{Exp}(1)$ reference; Appendix D.13 evaluates this diagnostic. Dataset-specific upper-tail departures motivate the EVT refinement below.

Step 3: Fit EVT to the reference tail.

Let

$$u = Q_{q_u}(\mathcal{Z}_{\mathrm{ref}})$$

denote the reference-tail anchor. The frozen confirmatory configuration uses $q_u = 0.95$. CAVET-AD fits a GPD to the strictly positive reference exceedances

$$\{Z - u : Z \in \mathcal{Z}_{\mathrm{ref}},\ Z > u\}$$

to obtain the shape and scale parameters (ξ, σ), following standard EVT tail modeling (Smith, 1987; Gumbel, 1958; Coles, 2001).

Step 4: Compute the tail-excess response.

Let u denote the frozen EVT anchor and let

$$p_u = \widehat{P}_{\mathrm{ref}}(Z > u)$$

denote the empirical reference tail mass above that anchor.

For an incoming batch of m observations, define the observed tail occupancy as

$$\hat{p}_{\mathrm{batch}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}(Z_i > u) \, .$$

CAVET-AD then defines the observable excess-tail response as

$$r_{\mathrm{tail}} = [\hat{p}_{\mathrm{batch}} - p_u]_+ \, ,$$

where $[a]_+ = \max(a, 0)$.

Interpretation. r_{tail} measures excess upper-tail mass relative to the calibration reference. It is not assumed to be a generally consistent estimator of the true contamination prevalence π.

The complete reference-calibration and tail-response procedure is summarized in Appendix B (Algorithm 1).

3.3 Tail-Excess Adaptive Threshold Scheduling

Given the EVT-refined reference survivor and an observed tail-excess response r_{tail}, CAVET-AD defines an adaptive target tail budget

$$b(r_{\mathrm{tail}}) = \min(b_{\max}, b_0 + \kappa r_{\mathrm{tail}}) \, ,$$

where b_0 is the nominal tail budget, κ controls adaptation strength, and $b_{\max}$ limits the maximum operating budget. For the frozen confirmatory configuration, these parameters were fixed at $b_0 = 0.01$, $\kappa = 0.20$, and $b_{\max} = 0.10$ and were not modified during confirmatory evaluation.

Let $p_u = \widehat{P}_{\mathrm{ref}}(Z > u)$ denote the empirical reference tail mass at the EVT anchor. We define the reference survivor used for threshold inversion as

$$\bar{F}_{\mathrm{ref,EVT}}(t) = \begin{cases} \widehat{\bar{F}}_{\mathrm{ref}}(t), & t \le u, \\ p_u \left(1 + \dfrac{\xi(t-u)}{\sigma}\right)^{-1/\xi}, & t > u, \end{cases}$$

where $\widehat{\bar{F}}_{\mathrm{ref}}(t)$ denotes the empirical survivor function of the reference Z scores. When $\xi \to 0$, the GPD tail uses its exponential limit. The adaptive CAVET-AD threshold is obtained by survivor inversion:

$$\tau^*(r_{\mathrm{tail}}) = \bar{F}^{-1}_{\mathrm{ref,EVT}}(b(r_{\mathrm{tail}})) \, .$$

The * superscript denotes the threshold selected by the adaptive operating rule; it does not imply Bayes or minimax optimality. Threshold inversion uses the empirical reference survivor when the requested budget places the operating point in the reference body and the GPD continuation when it lies beyond the EVT anchor. The adaptive operating rule is bounded and monotone in r_{tail}, and its finite-batch budget deviation admits an $O(m^{-1/2})$ high-probability bound under i.i.d. deployment-window sampling (Appendix C).

Adaptation-decision property. When $\hat{p}_{\text{batch}} \leq p_u$, we have $r_{\text{tail}} = 0$ and the threshold defaults to b_0; adaptation activates only when excess tail mass is present. This batch-time operating rule requires neither deployment labels nor knowledge of the true deployment prevalence.

The complete decision procedure is given in Appendix B (Algorithm 2).

3.4 Evaluation Principle

For each paired comparison, CAVET-AD and the fixed reference-tail rule receive the same frozen anomaly scores and evaluation mixture. Labels are used only by the harness to construct controlled prevalence mixtures and compute metrics (Section 4).

4 Datasets and Experimental Setup

4.1 Datasets

We evaluate CAVET-AD on four heterogeneous tabular benchmarks: CSE-CIC-IDS2018 (hereafter CICIDS2018) (Canadian Institute for Cybersecurity, University of New Brunswick, 2018), CICIoT2023 (Neto et al., 2023; Canadian Institute for Cybersecurity, University of New Brunswick, 2023), CreditCard2013 (Dal Pozzolo et al., 2018), and NSL-KDD (Tavallaee et al., 2009) (Table 16 in Appendix E). They span rare-anomaly financial transactions, moderate-prevalence intrusion detection, and a majority-anomaly IoT setting. For NSL-KDD, the canonical dataset combines `KDDTrain+.txt` and `KDDTest+.txt`; `KDDTrain+_20Percent.txt` is excluded because it is a subset of the full training set. Dataset statistics, natural prevalences, and construction details are reported in Appendix E.

All four benchmarks are tabular, isolating prevalence shift from modality-specific confounds. The CDF-based pipeline applies in principle to any continuous-score detector; we discuss deep-detector applicability and image-domain considerations in Section 5.8.

4.2 Frozen Confirmatory Protocol

For each dataset and seed, observations are partitioned into mutually exclusive *row-index* detector-training, reference-calibration, and held-out evaluation subsets. In this controlled protocol, the detector-training and reference-calibration partitions contain nominal observations only. Ground-truth labels are used only to construct these experimental partitions and controlled evaluation mixtures; they are not provided to the deployed adaptation rule. The detector is fitted only on the training partition, the reference partition constructs the empirical CDF and EVT tail model, and controlled prevalence mixtures are generated only from the evaluation pool. A post-confirmatory content audit identified repeated feature vectors across some row-disjoint partitions, so a stricter group-disjoint reconstruction was additionally evaluated without modifying the frozen primary results.

The confirmatory configuration uses ten seeds $\{7, 11, 13, 21, 23, 37, 51, 55, 73, 97\}$ and prevalence levels $\pi \in \{0, .01, .05, .10, .20, .30, .50, .60, .80\}$. The frozen operating parameters are $q_u = .95$, $b_0 = .01$, $\kappa = .20$, and $b_{\max} = .10$. The development seed 42 is excluded. Full partition provenance and frozen configuration details are reported in Tables 3 and 4.

4.3 Evaluation and Statistical Protocol

F1, precision, recall, and predicted-positive rate evaluate binary operating behavior; ROC–AUC and average precision (AP) are retained only as score-ranking diagnostics. At each dataset–seed–prevalence condition, CAVET-AD and the fixed reference-tail baseline receive the same realized evaluation mixture, yielding the paired effect

$$\Delta F_1 = F_{1,\text{CAVET}} - F_{1,\text{fixed}}.$$

Evaluation observations are sampled without replacement when possible, with mixture size reduced rather than duplicating rare anomalies. Seed-level paired effects are used for inference. The dataset-level descriptive summaries in Table 1 average all nine prespecified conditions, including the $\pi = 0$ all-normal sanity condition. Prevalence-resolved effectiveness inference is conducted separately

over the eight prespecified nonzero prevalence conditions; $\pi = 0$ is not treated as an effectiveness comparison.

Confidence intervals summarize uncertainty across the ten paired seed-level effects rather than individual evaluation observations. For every reported comparison, we use the two-sided 95% Student-t interval $\bar{\Delta} \pm t_{0.975,9} s_\Delta / \sqrt{10}$. A small-sample coverage audit used 3,000 Monte Carlo samples per scenario at $n = 10$, comparing percentile-bootstrap, BCa, and Student-t intervals under Gaussian, scaled t_5, centered Beta$(2, 5)$, and zero-inflated Gaussian effects. Empirical coverage ranged from 0.8770 to 0.8973 for percentile bootstrap, from 0.8210 to 0.8997 for BCa, and from 0.9450 to 0.9653 for Student-t; only Student-t satisfied the prespecified $[0.925, 0.975]$ acceptance band in all four scenarios. The intervals are unadjusted descriptive summaries. Multiplicity-adjusted conclusions and the "Sig." classifications use exact paired sign-flip tests. Holm correction is applied across the four prespecified dataset hypotheses for aggregate results and separately within each dataset across its eight prespecified nonzero-prevalence hypotheses for prevalence-resolved results. Interval inclusion of zero is not used to define those classifications.

The *relative F1 gain* in Table 1 is descriptive; statistical conclusions use the paired tests and prevalence-resolved results instead. Positive-class precision and recall accompany F1 in the primary and principal operating-point comparisons, using companion tables where including all quantities together would reduce readability.

Full-protocol replication and post-confirmatory analyses. After freezing the Isolation Forest experiment, we completed an ECOD replication over all $4 \times 10 \times 9 = 360$ paired dataset–seed–prevalence conditions (720 method-level rows), using the same partitions, prevalence mixtures, EVT anchor, budget rule, and statistical unit. Post-confirmatory analyses covered group-disjoint robustness, alternative thresholds, ablations, sensitivity, tail-identifiability, LOF transfer, the r_{tail}–gain association, natural-prevalence CreditCard2013, and controlled covariate shift.

The natural-prevalence experiment used the untouched CreditCard2013 pool at $\pi = 0.00173$, with 100,000 observations per seed across the ten frozen seeds. The covariate-shift test fixed prevalence at $\pi = 0.10$ and translated deployment features along the leading training-PCA direction at four anchor levels $\{0, 0.25, 0.50, 1.00\}$. To examine accumulated displacement, we also replayed the same seed-specific evaluation membership through 21 consecutive synthetic batches with $\delta_b = 0.05(b-1)$, $b = 1, \ldots, 21$. The detector, reference model, prevalence, evaluation membership, and frozen CAVET-AD configuration were not updated across batches.

We additionally compare against SPOT, a sequential EVT baseline, on the same frozen detector scores and exactly matched dataset–seed–prevalence mixtures. SPOT uses its returned alarm indices as the authoritative predictions; its returned thresholds are retained only as diagnostics and are not retrospectively reapplied. Because F1 is uninformative for the $\pi = 0$ all-normal condition, the primary SPOT comparison uses the 320 positive-prevalence conditions. Complete implementation and inferential results are reported in Appendix D.5.

As a direct conformal comparator, we convert the frozen oriented detector scores from the post-confirmatory group-disjoint evaluation into upper-tail conformal p-values using the corresponding nominal reference scores and apply Benjamini–Hochberg at the prespecified level $q = 0.05$ within each deployment batch (Bates et al., 2023). This Conformal–BH rule uses neither deployment labels nor true deployment prevalence and is evaluated on the same group-disjoint dataset–seed–prevalence realizations as CAVET-AD and the alternative reference-only threshold comparators. Full protocol and results are reported in Appendix D.6.

5 Results and Discussion

This section reports the frozen paired confirmatory evaluation of CAVET-AD against the fixed reference-tail rule. The primary effect is $\Delta F_1 = F_{1,\text{CAVET}} - F_{1,\text{fixed}}$, where positive values favor CAVET-AD. The dataset-level summaries in Table 1 cover the complete nine-condition sweep. Statistical effectiveness conclusions follow the paired nonzero-prevalence protocol in Section 4.3; $\pi = 0$ is interpreted only as an all-normal sanity condition.

Table 1: Primary frozen confirmatory results over the complete nine-condition sweep, including the $\pi = 0$ all-normal sanity condition (ten seeds). Mean ΔF_1 is the paired CAVET-AD minus fixed-rule F1 difference. CI denotes the two-sided seed-level Student-t interval, and Holm p denotes the multiplicity-adjusted sign-flip result. Prevalence-resolved effectiveness inference excludes $\pi = 0$ and is reported in Table 12; the corresponding positive-prevalence precision, recall, and F1 results are reported in Table 11. Tail ID denotes the retrospective tail-identifiability classification (Appendix D.10).

Dataset	Mean ΔF_1	95% CI	Holm p	Relative F1 gain	First CAVET-favorable π	Last fixed-favorable π	Tail ID
CreditCard2013	0.0904	[0.0743, 0.1066]	0.0078	13.9%	0.05	–	Positive
CICIoT2023	0.2683	[0.1957, 0.3410]	0.0078	88.9%	0.01	–	Positive
NSL-KDD	0.0107	[0.0075, 0.0138]	0.0078	1.1%	0.30	0.10	Positive
CICIDS2018	0.0000	[0.0000, 0.0000]	1.0000	0.0%	–	–	Negative

5.1 Primary Confirmatory Findings

The central result is that the operating-point benefit of CAVET-AD is *dataset- and prevalence-dependent*, rather than uniformly positive across all operating regimes.

Table 1 summarizes the aggregate seed-level confirmatory evidence. CICIoT2023 exhibits the largest overall benefit, followed by CreditCard2013. NSL-KDD exhibits a substantially smaller positive aggregate effect together with a clear prevalence-dependent transition. CICIDS2018 forms an informative no-effect boundary case.

These aggregate paired effects are statistically supported across the ten confirmatory seeds for CICIoT2023, CreditCard2013, and NSL-KDD after Holm correction. However, the aggregate result alone does not characterize the full operating behavior. In particular, CreditCard2013 exhibits a slightly negative directional effect at the lowest nonzero prevalence ($\pi = 0.01$) that does not remain significant after Holm correction, while NSL-KDD transitions from fixed-baseline favorability at low prevalence to CAVET-AD favorability at higher prevalence. The complete prevalence-resolved inference is reported in Table 12.

5.2 Prevalence-Resolved Operating-Point Behavior

The four datasets exhibit distinct prevalence-dependent regimes rather than a common monotone pattern (Figure 2). Paired estimates, Student-t intervals, Holm-adjusted sign-flip results, and inferential classifications are reported in Appendix D.8, Table 12.

Dataset-specific regimes. The prevalence-resolved results reveal four distinct deployment regimes. CICIoT2023 (majority-anomaly) shows the strongest adaptive benefit ($\overline{\Delta F_1} = 0.2683$, 95% CI $[0.1957, 0.3410]$, Holm $p = 0.0078$), with CAVET-AD favored from $\pi = 0.01$ onward because the extreme prevalence shift severely miscalibrates the fixed threshold. CreditCard2013 (rare-anomaly) also shows an aggregate benefit over the controlled sweep ($\overline{\Delta F_1} = 0.0904$, Holm $p = 0.0078$), but only above its native rare-event regime: the effect is slightly negative and not Holm-significant at $\pi = 0.01$, becomes significantly favorable from $\pi = 0.05$, and yields no multiplicity-adjusted significant F1 improvement at the natural prevalence $\pi = 0.00173$ (Appendix H.2).

NSL-KDD exhibits a prevalence-dependent crossover: the fixed rule is favored at $\pi \leq 0.10$, the result is unresolved at $\pi = 0.20$, and CAVET-AD is favored from $\pi = 0.30$ onward ($\overline{\Delta F_1} = 0.0107$). CICIDS2018 provides the informative boundary case: the frozen primary sweep yields exactly zero CAVET-versus-fixed difference, while the label-free r_{tail} remains near zero across all prevalence conditions. Thus, the method correctly retains the fixed behavior when the score distribution does not warrant adaptation (Section 5.4). Complete prevalence-specific inference and tail-identifiability results appear in Appendices D.8 and D.10.

5.3 Supporting Robustness and Baseline Evidence

Post-confirmatory analyses preserve the conditional interpretation of the primary result. The four operating regimes persist under group-disjoint reconstruction; comparisons with reference-only, sequential EVT, and conformal baselines reveal no uniformly dominant alternative; and ablations

identify adaptive budgeting as the principal source of improvement. At the natural CreditCard2013 prevalence, however, adaptation remains minimal and does not significantly improve F1 after multiplicity correction. Complete results, including paired statistics and sensitivity analyses, are reported in Appendix D.2.

5.4 When Does Adaptation Help? r_{tail} as a Deployment-Time Diagnostic

We test whether the label-free r_{tail} is associated with the adaptation benefit $\Delta F_{1,\mathrm{adapt}} = F_{1,\mathrm{full}} - F_{1,\text{no adapt}}$. Within-seed Spearman correlations are strongly positive for CreditCard2013 and CICIoT2023 (median $\rho = 1.0$) and NSL-KDD ($\rho \approx 0.93$); for CICIDS2018, r_{tail} remains near zero and the correlation is undefined. These results support r_{tail} as an applicability diagnostic under the evaluated pure prevalence-shift conditions: larger values are associated with greater adaptation benefit, whereas near-zero values correspond to little or no adaptation. A retrospective label-based analysis (Appendix D.10) provides complementary evidence that the observable response reflects tail identifiability.

5.5 Full-Protocol Replication with ECOD

We repeated the complete confirmatory protocol with ECOD (Li et al., 2023), retaining all four datasets, ten seeds, nine prevalence conditions, frozen partitions, and $(q, b_0, \kappa, b_{\max}) = (0.95, 0.01, 0.20, 0.10)$. ECOD was fitted only on the detector-training partition, while reference-CDF and EVT estimation remained restricted to the nominal reference partition.

The frozen calibration pipeline was operational with ECOD, but its benefit remained dataset-dependent (Table 23). CreditCard2013 improved from 0.7439 to 0.8055 ($\Delta F_1 = 0.0615$, 95% CI $[0.0588, 0.0643]$, Holm $p = 0.0078$ across the four prespecified ECOD dataset comparisons). CICIDS2018 showed a small non-significant difference, while CICIoT2023 and NSL-KDD showed no observed difference. Thus, ECOD supports portability of the frozen calibration procedure, not universal improvement across detectors or datasets. Complete protocol and supplementary LOF results are reported in Appendix G.6.

A detector-ranking audit clarifies the CICIDS2018 boundary. Isolation Forest is near chance (ROC–AUC 0.5162), ECOD is below chance (ROC–AUC 0.3422), whereas LOF remains strongly informative (ROC–AUC 0.9405) and supports a substantial adaptive operating-point gain. Thus, the CICIDS2018 outcome depends on the informativeness of the underlying detector scores rather than on threshold adaptation alone (Appendix G.6).

5.6 Held-Out Evaluation of the Applicability Signal

The contemporaneous association between r_{tail} and adaptation benefit does not by itself establish whether the signal transfers to an unseen dataset. We therefore conducted a leave-one-dataset-out policy evaluation: in each fold, the r_{tail} decision cutoff was fitted using only the other three datasets and frozen before evaluation on the held-out dataset.

Across 40 held-out dataset–seed units, the resulting policy improved mean F_1 over the fixed rule by 0.0982 (two-sided Wilcoxon $p = 2.14 \times 10^{-7}$). Its mean F_1 was only 0.0057 below always adapting, a difference that was not significant ($p = 0.131$), and it selected the oracle action in 83.1% of 320 conditions. On CICIDS2018, it retained the fixed rule in every condition, demonstrating the intended no-adaptation behavior. This is a cross-fitted evaluation rather than a prospective temporal deployment. The complete protocol, leakage audit, dataset-level results, and table are reported in Appendix D.11.

5.7 Controlled Covariate-Shift Stress Test

To test whether excess tail mass is specific to prevalence change, we translated CreditCard2013 deployment features along the leading training-PCA direction at shifts $\{0, 0.25, 0.50, 1.00\}$, holding prevalence ($\pi = 0.10$), the detector, reference model, partitions, and CAVET-AD configuration fixed across ten paired seeds. CAVET-AD remained significantly favorable at shifts 0 and 0.25 (mean $\Delta F_1 = 0.0642$ and 0.0459), was slightly and not Holm-significantly unfavorable at 0.50 ($\Delta F_1 = -0.0089$), and deteriorated sharply at 1.00 even as r_{tail} rose to 0.3390 ($\Delta F_1 = -0.5102$,

Holm $p = 0.0078$). Thus, tail response alone does not identify beneficial adaptation under non-prevalence shift. Appendix G.5 reports the full inference, and Appendix G.9 provides a study-specific dual-criterion KS safeguard with a frozen-seed example.

In the complementary 21-batch gradual replay, the safeguard first activated at median shift 0.65 (range 0.55–0.70), was active for $8/10$ seeds at 0.65 and all seeds from 0.70 onward, and remained active thereafter (Figure 8).

5.8 Discussion and Deployment Interpretation

CAVET-AD is conditional: when $r_{\mathrm{tail}} \approx 0$, as on CICIDS2018, the budget remains near b_0 and recovers the fixed rule; under prevalence shift, larger responses may warrant adaptation. The leave-one-dataset-out result supports held-out transfer but is cross-fitted, not prospective.

Before deployment, practitioners should audit score ranking, independent nominal calibration, EVT-tail support, window-size stability, and non-prevalence drift; Appendix G.9 gives a concrete pre-check and fallback, though the score-ranking audit may draw on a small historical labeled sample: *label-free* here refers to the deployed adaptation rule itself, not end-to-end detector qualification. Because a large r_{tail} under other shifts does not itself justify adaptation, applications requiring batch-independent decisions should retain a fixed rule. The LOF sign reversal (favorable on CICIDS2018/NSL-KDD, unfavorable on CreditCard2013/CICIoT2023; Appendix G.6) has no validated ex-ante predictor and is a post-hoc explanation, not a deployment-selection rule. The frozen $(q, b_0, \kappa, b_{\max}) = (0.95, 0.01, 0.20, 0.10)$ configuration and heterogeneous ECOD/LOF results preclude universal detector-level claims; Appendix G provides further considerations.

5.9 Limitations

Evidence covers four tabular benchmarks, mainly shallow detectors; the natural-prevalence and controlled covariate-shift studies focus on CreditCard2013. CAVET-AD assumes representative nominal reference data and uses a heuristic budget plus an applicability cutoff fitted from labeled development outcomes, so benefits remain detector-, dataset-, prevalence-, and shift-dependent. With ten nonzero pairs, exact two-sided sign-flip inference has raw floor $2/2^{10} = 0.001953125$ and best-case eight-condition within-dataset Holm floor 0.015625. We added no seeds after inspecting the frozen outcomes; finer resolution requires a separately prespecified larger-seed replication. Appendix D.12 provides the complete limitations and validity assessment.

6 Conclusion

CAVET-AD addresses a central deployment question in unsupervised anomaly detection: when does an observed score distribution provide sufficient evidence to adapt the decision threshold? Its excess-tail response r_{tail} controls a bounded adaptation budget without deployment labels, known prevalence, or detector retraining. Across 320 positive-prevalence conditions, CAVET-AD increases descriptive mean F1 from 0.475 to 0.579. The benefit is strongly regime dependent: adaptation helps when prevalence change produces identifiable excess tail mass, may become preferable only after a prevalence-dependent crossover, and naturally returns toward the fixed rule when the response remains near zero.

A leave-one-dataset-out policy retains most of the adaptive benefit on a benchmark excluded from cutoff selection. ECOD replication, LOF transfer, alternative-threshold comparisons, and ablations provide additional evidence that the framework can transfer across detectors, while also showing that its benefit remains detector-, dataset-, and prevalence-dependent. At the natural CreditCard2013 prevalence, adaptation is minimal and yields no statistically significant F1 improvement after multiplicity correction. Under severe controlled covariate shift, r_{tail} can instead increase when adaptation is harmful. These results define the intended scope of CAVET-AD: it is a diagnostic operating rule for deciding when prevalence-driven threshold adaptation is warranted, not a general detector of distribution shift.

Future work should extend this decision framework to natural temporal and combined shifts, deep and non-tabular detectors, direct prevalence-estimation comparisons, adaptive conformal calibration, and data-driven budget schedules that do not depend on labeled deployment outcomes.

Ethics Statement

This work studies threshold calibration for unsupervised anomaly-detection systems, including applications in cybersecurity, fraud detection, and operational monitoring. Improved calibration may reduce missed anomalies or excessive false alarms; however, automated anomaly decisions may affect individuals or operational processes. CAVET-AD does not eliminate biases or limitations of the underlying detector or data distribution. High-stakes deployment should therefore include domain-specific validation, monitoring, and appropriate human oversight.

Reproducibility Statement

The experimental protocol, frozen configuration, dataset construction, partitioning, statistical procedures, and primary evaluation are documented in Sections 3–5. The appendices provide the complete algorithm, partition provenance, prevalence-resolved results, baseline comparisons, ablations, sensitivity analyses, group-disjoint robustness, detector-transfer results, and additional reproducibility details.

AI Use Statement

Generative AI tools were used to provide feedback on research methodology and experimental design, assist with software development and debugging, support interpretation and consistency checking of experimental results, identify and summarize relevant literature, and assist with manuscript drafting, editing, organization, and reference formatting. All AI-assisted content, code, experimental results, analyses, and citations were reviewed and verified by the authors against the underlying implementation, experimental outputs, and primary sources. The authors take responsibility for the final content of this work.

References

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

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104. Association for Computing Machinery, 2000. doi: 10.1145/342009.335388.

Canadian Institute for Cybersecurity, University of New Brunswick. CSE-CIC-IDS2018 on AWS. `https://www.unb.ca/cic/datasets/ids-2018.html`, 2018. Accessed: 2026-09-22.

Canadian Institute for Cybersecurity, University of New Brunswick. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. `https://www.unb.ca/cic/datasets/iotdataset-2023.html`, 2023. Accessed: 2026-09-22.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001. doi: 10.1007/978-1-4471-3675-0.

Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3784–3797, 2018. doi: 10.1109/TNNLS.2017.2736643.

Neema Davis, Gaurav Raina, and Krishna Jagannathan. A framework for end-to-end deep learning-based anomaly detection in transportation networks. *Transportation Research Interdisciplinary Perspectives*, 5:100112, 2020. doi: 10.1016/j.trip.2020.100112.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024. URL https://jmlr.org/papers/v25/22-1218.html.

Patrick Godau, Piotr Kalinowski, Evangelia Christodoulou, Annika Reinke, Minu Tizabi, Luciana Ferrer, Paul Jäger, and Lena Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment. *Medical Image Analysis*, 102:103504, 2025. doi: 10.1016/j.media.2025.103504.

Emil Julius Gumbel. *Statistics of Extremes*. Columbia University Press, 1958. doi: 10.7312/gumb92958.

Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2685–2693, 2016.

Jia Li, Shimin Di, Yanyan Shen, and Lei Chen. FluxEV: A fast and effective unsupervised framework for time-series anomaly detection. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 824–832. Association for Computing Machinery, 2021. doi: 10.1145/3437963.3441823.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2023. doi: 10.1109/TKDE.2022.3159580.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE Computer Society, 2008. doi: 10.1109/ICDM.2008.17.

Natalia Martinez Gil, Fearghal O'Donncha, Wesley Gifford, Nianjun Zhou, Dhaval Patel, and Roman Vaculin. Adaptive conformal anomaly detection with time series foundation models for signal monitoring. In *International Conference on Learning Representations*, 2026.

Euclides Carlos Pinto Neto, Sajjad Dadkhah, Raphael Ferreira, Alireza Zohourian, Rongxing Lu, and Ali A. Ghorbani. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors*, 23(13):5941, 2023. doi: 10.3390/s23135941.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. doi: 10.1214/aoms/1177704472.

Lorenzo Perini, Vincent Vercruyssen, and Jesse Davis. Transferring the contamination factor between anomaly detection domains by shape similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4128–4136, 2022. doi: 10.1609/aaai.v36i4.20331.

Lorenzo Perini, Paul-Christian Bürkner, and Arto Klami. Estimating the contamination factor's distribution in unsupervised anomaly detection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27668–27679. PMLR, 2023.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. doi: 10.1162/089976601750264965.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 838–846. PMLR, 2015.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1067–1075. Association for Computing Machinery, 2017. doi: 10.1145/3097983.3098144.

Richard L. Smith. Estimating tails of probability distributions. *The Annals of Statistics*, 15(3): 1174–1207, 1987. doi: 10.1214/aos/1176350499.

Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6. IEEE, 2009. doi: 10.1109/CISDA.2009.5356528.

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

Shuai Zhang, Chuan Zhou, Yang Liu, Peng Zhang, Xixun Lin, and Shirui Pan. Conformal anomaly detection in event sequences. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77177–77193. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25dn.html.

Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674. Association for Computing Machinery, 2017. doi: 10.1145/3097983.3098052.

A Positioning Relative to Prior Work

Table 2 provides a detailed comparison of CAVET-AD with representative anomaly-detection, calibration, and thresholding approaches. It complements the condensed discussion in Section 2 by separating the operating assumptions and deployment mechanisms of the compared approaches.

Table 2: Positioning of CAVET-AD relative to representative anomaly-detection and threshold-calibration approaches relevant to prevalence shift.

Research Area	Key Approaches	Core Mechanism	Limitation in Prior Work	CAVET-AD (This Work)
Score-based AD	IF, LOF, OCSVM, KDE, GMM, ECOD (Liu et al., 2008; Breunig et al., 2000; Schölkopf et al., 2001; Parzen, 1962; Bishop, 2006; Li et al., 2023; Ruff et al., 2021)	Produces anomaly, density, or decision scores	Requires an operating rule to convert continuous scores into anomaly decisions	Reference-tail calibration with a prevalence-responsive adaptive operating rule
EVT-based thresholding	SPOT/DSPOT, FluxEV, DeepEVT (Siffer et al., 2017; Davis et al., 2020; Li et al., 2021)	GPD-based modeling of extreme-score tails	Requires tail-region and operating-level specification; a fixed operating point does not itself adapt to prevalence shift	Applies EVT in calibrated Z-space with adaptive survivor inversion driven by an unsupervised tail response
Conformal AD	CP, online CP, adaptive conformal AD (Vovk et al., 2005; Angelopoulos & Bates, 2023; Gibbs & Candès, 2024; Zhang et al., 2025; Martinez Gil et al., 2026)	Distribution-free calibration and sequential/adaptive uncertainty control	Guarantees depend on exchangeability or the adaptation setting; requires a chosen rejection level	Uses reference-CDF calibration and EVT-survivor inversion without detector retraining
Contamination / prevalence estimation	CDE, AlphaMax, PU-learning, TRADE (Scott, 2015; Jain et al., 2016; Perini et al., 2022; 2023)	Estimates contamination or class prevalence from unlabeled deployment data	Requires explicit estimation of the deployment mixture proportion and corresponding method-specific assumptions	Uses observable tail-excess response in calibrated score space without directly estimating deployment prevalence
Contamination-aware AD	Contamination-factor modeling (Perini et al., 2023)	Models uncertainty in the contamination factor	Characterizes uncertainty in the contamination factor rather than adapting the deployment operating point from observed tail response	Adapts the operating threshold from observable reference-tail response without directly estimating deployment prevalence

A.1 Detailed Literature Context

Unsupervised anomaly detectors such as Isolation Forest, LOF, One-Class SVM, ECOD, KDE, Gaussian mixture models, and deep reconstruction methods produce continuous anomaly scores whose scales and interpretations differ across detectors (Liu et al., 2008; Breunig et al., 2000; Schölkopf et al., 2001; Li et al., 2023; Parzen, 1962; Bishop, 2006; Zhou & Paffenroth, 2017). Deployment therefore requires a separate operating rule that converts these scores into binary decisions. Fixed contamination assumptions, score quantiles, or validation operating points can become poorly calibrated when the deployment prevalence or score distribution changes.

EVT-based methods provide a principled approach to modeling extreme-score tails. SPOT/DSPOT use peaks-over-threshold and GPD modeling for streaming anomaly thresholds (Siffer et al., 2017), while subsequent work has extended EVT-based thresholding to additional anomaly-detection settings (Li et al., 2021; Davis et al., 2020). These methods establish the usefulness of tail modeling, but a deployment decision still requires specification of an operating risk or exceedance level. CAVET-AD instead uses EVT as a reference-tail refinement component while adapting the operating budget from an observable excess-tail response.

Conformal calibration. Conformal methods transform nonconformity scores into calibrated p-values or prediction sets under exchangeability assumptions (Vovk et al., 2005; Bates et al., 2023). Multiple conformal outlier hypotheses can subsequently be processed using false-discovery-rate control, while online and adaptive conformal methods extend calibration to sequential or shifting settings (Gibbs & Candès, 2024; Martinez Gil et al., 2026). CAVET-AD addresses a complementary operating-point problem: it adapts the alert budget from observable tail-mass changes rather than controlling a prespecified rejection or coverage level.

Prevalence estimation. Methods that estimate deployment prevalence or contamination (Scott, 2015; Jain et al., 2016; Perini et al., 2022; 2023) require identifiability assumptions such as known source contamination or component separability. CAVET-AD instead measures excess tail mass without recovering the true prevalence. These approaches are complementary.

B Additional Algorithms and Implementation Details

Algorithm 1 CAVET-AD: Reference Calibration and Tail-Excess Response

Input: Reference scores $\mathcal{S}_{\mathrm{ref}}$; incoming batch scores $\mathcal{S}_{\mathrm{batch}}$; detector score orientation; EVT threshold quantile q_u
Output: Reference CDF $\widehat{F}_{\mathrm{ref}}$, EVT parameters (u, ξ, σ), calibrated batch scores $\{Z_i\}$, and tail-excess response r_{tail}
Step 1: Reference CDF estimation
Fix detector score orientation so that larger values indicate stronger anomaly evidence
Set $N_{\mathrm{ref}} \leftarrow |\mathcal{S}_{\mathrm{ref}}|$
$\widehat{F}_{\mathrm{ref}}(s) \leftarrow \frac{\#\{s_j \in \mathcal{S}_{\mathrm{ref}} : s_j \le s\}}{N_{\mathrm{ref}}+1}$
Step 2: Reference and batch transformation
for $s \in \mathcal{S}_{\mathrm{ref}}$ **do**
    $Z_{\mathrm{ref}}(s) \leftarrow -\log\Big(1 - \widehat{F}_{\mathrm{ref}}(s)\Big)$
end for
for $s \in \mathcal{S}_{\mathrm{batch}}$ **do**
    $Z_{\mathrm{batch}}(s) \leftarrow -\log\Big(1 - \widehat{F}_{\mathrm{ref}}(s)\Big)$
end for
Step 3: EVT tail modeling on the reference distribution
Set $u \leftarrow Q_{q_u}(\mathcal{Z}_{\mathrm{ref}})$
Form reference exceedances $\{Z - u : Z \in \mathcal{Z}_{\mathrm{ref}}, Z > u\}$
Fit a Generalized Pareto Distribution to obtain (ξ, σ)
Step 4: Tail-excess response
$p_u \leftarrow \widehat{P}_{\mathrm{ref}}(Z > u)$
$\hat{p}_{\mathrm{batch}} \leftarrow \frac{1}{m}\sum_{Z \in \mathcal{Z}_{\mathrm{batch}}} \mathbf{1}(Z > u)$
$r_{\mathrm{tail}} \leftarrow [\hat{p}_{\mathrm{batch}} - p_u]_+$
return $\widehat{F}_{\mathrm{ref}}$, $\mathcal{Z}_{\mathrm{batch}}$, (u, ξ, σ), r_{tail}

Algorithm 2 CAVET-AD Adaptive Threshold and Decision Rule

Input: EVT-refined reference survivor $\bar{F}_{\mathrm{ref,EVT}}$; tail-excess response r_{tail}; base budget b_0; adaptation coefficient κ; maximum budget $b_{\max}$; batch scores $\{Z_i\}$
Output: Adaptive threshold τ^* and binary predictions $\{\hat{y}_i\}$
Compute the adaptive tail budget

$$b \leftarrow \min(b_{\max}, b_0 + \kappa r_{\mathrm{tail}})$$

Obtain the adaptive threshold by survivor inversion

$$\tau^* \leftarrow \bar{F}^{-1}_{\mathrm{ref,EVT}}(b)$$

for each calibrated score Z_i **do**
    $\hat{y}_i \leftarrow \mathbf{1}(Z_i \ge \tau^*)$
end for
return τ^* and $\{\hat{y}_i\}$

Algorithm 3 Frozen Paired Confirmatory Evaluation

Input: Datasets $\{D^m\}$; prespecified seed set S; prespecified prevalence grid Π
for each dataset D^m **do**
  for each seed $s \in S$ **do**
    Construct mutually exclusive row-index detector-training, reference, and evaluation partitions
    Fit the base detector using only the detector-training partition
    Construct the reference CDF and EVT model using only the reference partition
    for each controlled prevalence $\pi \in \Pi$ **do**
      Construct a controlled mixture only from the held-out evaluation pool
      Evaluate the frozen fixed reference-tail operating rule
      Evaluate CAVET-AD on the identical mixture
      Store $\Delta F_1 = F_{1,\mathrm{CAVET}} - F_{1,\mathrm{fixed}}$
    end for
  end for
end for
Perform seed-level paired inference only after all confirmatory results have been frozen

C Basic Properties of the Adaptive Operating Rule

The following proposition records deterministic properties of the CAVET-AD operating rule. These properties characterize the direction and boundedness of adaptation; they do not imply predictive optimality or improved F1.

Proposition 1 (Bounded monotone adaptation). Let

$$r_{\mathrm{tail}} = [\hat{p}_{\mathrm{batch}} - p_u]_+, \qquad b(r) = \min\{b_{\max},\, b_0 + \kappa r\},$$

where $0 < b_0 \le b_{\max} < 1$ and $\kappa \ge 0$. Let

$$\tau^*(r) = \bar{F}^{-1}_{\mathrm{ref,EVT}}(b(r)),$$

where $\bar{F}_{\mathrm{ref,EVT}}$ is a non-increasing reference survivor function. Then:

$$r_{\mathrm{tail}} \ge 0, \qquad b_0 \le b(r_{\mathrm{tail}}) \le b_{\max},$$

and, for any $0 \le r_1 \le r_2$,

$$b(r_1) \le b(r_2), \qquad \tau^*(r_1) \ge \tau^*(r_2).$$

If

$$r_{\mathrm{tail}} \ge \frac{b_{\max} - b_0}{\kappa}$$

for $\kappa > 0$, then $b(r_{\mathrm{tail}}) = b_{\max}$, so further increases in the observed tail response leave the operating threshold unchanged.

Proof. The positive-part definition gives $r_{\mathrm{tail}} \ge 0$. Because $\kappa \ge 0$, $b_0 + \kappa r$ is non-decreasing in r; taking its minimum with $b_{\max}$ preserves monotonicity and gives $b_0 \le b(r) \le b_{\max}$. A survivor function is non-increasing, so its generalized inverse is non-increasing in the requested survivor probability. Hence $b(r_1) \le b(r_2)$ implies $\tau^*(r_1) \ge \tau^*(r_2)$. Finally, when $r \ge (b_{\max} - b_0)/\kappa$, the uncapped budget is at least $b_{\max}$, and therefore the minimum equals $b_{\max}$. □

Proposition 2 (Finite-batch stability of the adaptive budget). Let $Z_1, \ldots, Z_m$ be i.i.d. calibrated scores from a fixed deployment-window distribution, and define

$$p_{\mathrm{batch}} = \Pr(Z > u), \qquad \hat{p}_{\mathrm{batch}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(Z_i > u).$$

Define the population and empirical excess-tail responses as

$$r^{\circ}_{\mathrm{tail}} = [p_{\mathrm{batch}} - p_u]_+, \qquad \hat{r}_{\mathrm{tail}} = [\hat{p}_{\mathrm{batch}} - p_u]_+.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|\hat{r}_{\mathrm{tail}} - r^{\circ}_{\mathrm{tail}}| \le \sqrt{\frac{\log(2/\delta)}{2m}}.$$

For

$$b(r) = \min\{b_{\max},\, b_0 + \kappa r\}, \qquad \kappa \ge 0,$$

the corresponding adaptive budgets satisfy, with the same probability,

$$|b(\hat{r}_{\mathrm{tail}}) - b(r^{\circ}_{\mathrm{tail}})| \le \kappa \sqrt{\frac{\log(2/\delta)}{2m}}.$$

Proof. Under the stated i.i.d. assumption, the variables $\mathbf{1}(Z_i > u)$ are independent Bernoulli random variables with mean p_{batch}. Hoeffding's inequality therefore gives

$$\Pr(|\hat{p}_{\mathrm{batch}} - p_{\mathrm{batch}}| \ge \epsilon) \le 2\exp(-2m\epsilon^2).$$

Setting

$$\epsilon = \sqrt{\frac{\log(2/\delta)}{2m}}$$

gives, with probability at least $1-\delta$,

$$|\hat{p}_{\mathrm{batch}} - p_{\mathrm{batch}}| \le \sqrt{\frac{\log(2/\delta)}{2m}}.$$

The positive-part map $x \mapsto [x - p_u]_+$ is 1-Lipschitz. Therefore,

$$|\hat{r}_{\mathrm{tail}} - r^{\circ}_{\mathrm{tail}}| \le |\hat{p}_{\mathrm{batch}} - p_{\mathrm{batch}}|,$$

which proves the first claim.

The clipping map

$$b(r) = \min\{b_{\max},\, b_0 + \kappa r\}$$

is κ-Lipschitz for $\kappa \ge 0$. Hence

$$|b(\hat{r}_{\mathrm{tail}}) - b(r^{\circ}_{\mathrm{tail}})| \le \kappa\, |\hat{r}_{\mathrm{tail}} - r^{\circ}_{\mathrm{tail}}|.$$

Combining this inequality with the preceding concentration bound yields

$$|b(\hat{r}_{\mathrm{tail}}) - b(r^{\circ}_{\mathrm{tail}})| \le \kappa \sqrt{\frac{\log(2/\delta)}{2m}},$$

with probability at least $1-\delta$. □

D Supplementary Statistical and Calibration Diagnostics

This appendix provides supplementary diagnostics for the frozen confirmatory experiment reported in Section 5. The primary statistical conclusions are based on the paired dataset–seed–prevalence comparisons and multiplicity-adjusted inference reported in the main paper.

The complete frozen confirmatory configuration is summarized in Table 3. The table is reconstructed directly from the frozen protocol and executed confirmatory result objects. All executed result tables agree on the ten-seed set $S = \{7, 11, 13, 21, 23, 37, 51, 55, 73, 97\}$ and the nine prespecified prevalence conditions. No parameter was selected or modified during this reconstruction.

Table 3: Frozen confirmatory design and operating configuration. The configuration was fixed before the multi-seed confirmatory evaluation. Detector fitting, reference calibration, and evaluation use mutually disjoint partitions. Ground-truth labels and true prevalence are not supplied to the deployed CAVET-AD operating rule.

Design component	**Frozen specification**
Datasets	CreditCard2013; CICIDS2018; CICIoT2023; NSL-KDD
Confirmatory seeds	$7, 11, 13, 21, 23, 37, 51, 55, 73, 97$
Controlled prevalence grid	$\{0.00, 0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.60, 0.80\}$
Partition design	Nominal detector-training, nominal reference, and held-out evaluation partitions
Partition separation	Detector-training, reference, and evaluation partitions are mutually disjoint by row index; feature-content separation is evaluated separately in the post-confirmatory group-disjoint robustness experiment
Base detector	Isolation Forest, 300 trees; detector fitted only on the detector-training partition
Preprocessing fit source	Detector-training partition only
Reference ECDF fit source	Frozen reference partition only
EVT anchor	$q = 0.95$
EVT model	Generalized Pareto tail model; frozen MLE fitting with fallback; fitted only on the reference partition
Adaptive-budget parameters	$b_0 = 0.01$, $\kappa = 0.20$, $b_{\max} = 0.10$
Adaptive-budget rule	$b = \min(b_{\max},\, b_0 + \kappa r_{\text{tail}})$
Ground-truth labels in deployed rule	No
True prevalence supplied to adaptation	No
Paired method evaluation	Yes; CAVET-AD and the fixed reference-tail rule are evaluated on the same dataset–seed–prevalence mixture and evaluation size

Table 4: Frozen row-index partition and provenance audit. Train and Ref. denote the detector-training and reference partitions, respectively; both contain nominal observations only. Eval. denotes the held-out evaluation pool. Eval. π is the natural anomaly prevalence of that untouched pool before construction of controlled-prevalence mixtures. The partitions are mutually exclusive by row index, and Train + Ref. + Eval. equals the processed dataset total exactly. Feature-content uniqueness is evaluated separately in the post-confirmatory group-disjoint robustness analysis.

Dataset	Total	Train	Ref.	Eval.	Eval. normal	Eval. anomaly	Eval. π
CreditCard2013	284,807	56,863	56,863	171,081	170,589	492	0.288%
CICIDS2018	13,522,722	2,382,409	2,382,409	8,757,904	7,147,231	1,610,673	18.391%
CICIoT2023	45,019,243	210,274	210,274	44,598,695	630,825	43,967,870	98.586%
NSL-KDD	148,517	15,410	15,410	117,697	46,234	71,463	60.718%

D.1 Frozen Row-Index Partition and Provenance Audit

The frozen confirmatory partition audit verifies the realized *row-index-disjoint* detector-training, reference, and held-out evaluation partitions for all four datasets. Detector-training and reference partitions contain only nominal observations, whereas anomalous observations are retained exclusively in the held-out evaluation pool for construction of controlled prevalence mixtures. For every dataset, the three partition counts sum exactly to the corresponding processed dataset size.

This audit establishes separation at the row-index and experimental-protocol level; it does not by itself guarantee feature-content uniqueness across partitions. A later post-confirmatory content audit identified repeated feature-content vectors in the canonical processed datasets. The stricter group-disjoint robustness experiment reported in Section D.3 assigns identical feature-content groups to only one partition and is analyzed separately from the frozen primary confirmatory experiment.

The audit additionally verifies that preprocessing and base-detector fitting use only the detector-training partition; the reference empirical CDF and EVT tail model are fitted only on the frozen reference partition; and the evaluation pool is excluded from all of these fitting stages. The deployed CAVET-AD adaptation rule uses neither ground-truth labels nor the true anomaly prevalence. These checks were performed from the frozen protocol and partition objects without recomputing any partition, preprocessing pipeline, detector, reference model, EVT fit, or confirmatory result.

D.2 Complete Supporting Robustness and Baseline Evidence

Post-confirmatory analyses preserve the conditional interpretation of the primary result. A group-disjoint reconstruction prevents identical feature-content groups from crossing partitions and retains the principal four-dataset regimes (Appendix D.3). Comparisons with empirical reference-$q = .99$, robust-MAD, and Tukey-IQR thresholds show that no reference-only rule uniformly dominates (Appendix D.4).

We further compare against SPOT, a sequential EVT baseline, on exactly matched dataset–seed–prevalence conditions. Across the 320 positive-prevalence conditions, CAVET-AD obtains a descriptive mean F1 of 0.579, compared with 0.463 for SPOT. CAVET-AD is significantly better on CICIoT2023, CreditCard2013, and NSL-KDD after Holm correction, whereas CICIDS2018 remains an indistinguishable no-benefit boundary case (Appendix D.5).

In the post-confirmatory group-disjoint evaluation, CAVET-AD obtains a descriptive mean F1 of 0.582, compared with 0.381 for a direct Conformal–BH comparator, across 320 exactly paired positive-prevalence conditions. The mean paired difference is $+0.201$; CAVET-AD has higher F1 in 299 conditions and lower F1 in 21. Dataset-level differences remain significant after Holm correction on all four datasets, although the absolute CICIDS2018 difference is only 0.0086 and reflects the near-absence of Conformal–BH rejections rather than a large practical gain (Appendix D.6).

Component ablation identifies adaptive budgeting as the principal source of improvement, whereas removing EVT refinement produces much smaller aggregate changes (Appendix D.7). At the natural CreditCard2013 prevalence of 0.173%, the adaptive budget remained close to the fixed budget and mean F1 changed only from 0.16765 to 0.16658 ($\overline{\Delta F_1} = -0.00107$; Holm $p = 0.108$). Thus, the favorable controlled-sweep result does not imply improvement at the dataset's native prevalence (Appendix H.2). Finally, EVT-anchor, budget, EVT-fitting, and nonlinear-schedule analyses show dataset-dependent sensitivity but no alternative schedule that consistently outperforms the frozen linear rule (Appendix G).

D.3 Group-Disjoint Robustness Evaluation

The frozen confirmatory experiment used mutually exclusive row-index partitions. A post-confirmatory audit subsequently identified repeated feature-content vectors across partitions in the canonical datasets. To test whether such duplication materially affected the conclusions, we constructed a stricter group-disjoint robustness evaluation in which identical feature-content groups were assigned to only one partition.

For each dataset and confirmatory seed, pure-normal duplicate groups were assigned wholly to detector-training, reference, or evaluation partitions, while mixed-label groups were conservatively restricted to evaluation. Anomalous observations remained evaluation-only. All 40 dataset–seed combinations passed the resulting partition-integrity checks, including nominal-only detector-training and reference partitions and group-level separation of repeated feature content.

Re-running the Fixed-versus-CAVET comparison under these stricter partitions preserved the substantive conclusions of the frozen confirmatory analysis. CAVET retained strong adaptive gains on CICIoT2023, remained favorable in aggregate on CreditCard2013, preserved the prevalence-dependent crossover on NSL-KDD, and continued to show negligible practical separation from the fixed rule on CICIDS2018. Thus, the principal conclusions are not explained by cross-partition duplication of identical feature content.

This analysis is a post-confirmatory robustness check and does not replace or modify the original frozen row-index-disjoint confirmatory experiment.

Table 5: Prevalence-resolved robustness of the CAVET-AD versus fixed-reference-tail comparison under the stricter group-disjoint partition construction. Entries report the mean paired $\Delta F_1 = F_{1,\mathrm{CAVET}} - F_{1,\mathrm{fixed}}$ across the ten confirmatory seeds. "Original" denotes the frozen row-index-disjoint experiment and "Group" denotes the post-confirmatory group-disjoint robustness experiment.

Dataset	$\pi = 0.01$		$\pi = 0.10$		$\pi = 0.30$		$\pi = 0.80$	
	Original	Group	Original	Group	Original	Group	Original	Group
CreditCard2013	-0.0073	-0.0073	0.0687	0.0675	0.1213	0.1162	0.1825	0.1741
CICIDS2018	0.0000	-4.69×10^{-7}	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CICIoT2023	0.0158	0.0343	0.2810	0.3120	0.3681	0.3766	0.4263	0.4303
NSL-KDD	-0.0272	-0.0241	-0.0315	-0.0338	0.0152	0.0111	0.0781	0.0762

Table 5 shows that the principal prevalence-dependent operating regimes remain qualitatively stable after eliminating cross-partition duplication of identical feature content. CICIoT2023 remains strongly CAVET-favorable, CreditCard2013 retains its lowest-prevalence qualification followed by positive adaptive gains, and NSL-KDD retains the low-prevalence fixed-favorable to higher-prevalence CAVET-favorable crossover.

CICIDS2018 requires a more precise interpretation. The frozen primary experiment produced exactly zero CAVET-versus-Fixed difference throughout the prevalence sweep. Under group-disjoint splitting, 89 of the 90 dataset–seed–prevalence realizations remain exactly unchanged. The only nonzero change occurs at seed 13 and $\pi = 0.01$, yielding a prevalence-level mean difference of approximately -4.69×10^{-7} and a maximum seed-level absolute change of approximately 4.69×10^{-6}. Accordingly, we report the numerical discrepancy explicitly rather than reclassifying it using an arbitrary practical-significance tolerance. The experiment therefore preserves the substantive no-benefit boundary interpretation of CICIDS2018 while avoiding a claim of exact numerical identity under the stricter partition construction.

D.4 Alternative Reference-Only Threshold Comparison

The post-confirmatory comparator study evaluates five operating rules on the same group-disjoint dataset–seed–prevalence realizations: CAVET-AD, the fixed reference-tail rule, empirical reference $q = 0.99$, robust MAD, and Tukey-IQR.

All comparator thresholds use reference information only. The empirical $q = 0.99$, robust-MAD, and Tukey-IQR thresholds are computed on oriented raw detector scores, whereas CAVET-AD and the fixed reference-tail rule operate in the calibrated reference-tail representation.

The comparison remains regime-dependent. CAVET-AD has the highest aggregate mean F1 on CICIoT2023, CreditCard2013, and NSL-KDD, whereas Tukey-IQR has the highest aggregate mean F1 on CICIDS2018. These results reinforce the main-text conclusion that no single thresholding rule uniformly dominates across all datasets and prevalence regimes.

D.5 Comparison with Sequential EVT Calibration

SPOT provides a sequential peaks-over-threshold comparator that updates its tail model as deployment scores arrive. We evaluate SPOT using the same frozen detector-score orientation and exactly matched dataset–seed–prevalence mixtures used by the primary CAVET-AD evaluation. The SPOT implementation uses $q = 0.01$, initial level 0.98, ten initial points, and alarm reporting enabled. These parameters were fixed before examining comparative performance.

The alarm indices returned by SPOT are treated as the authoritative binary predictions. Returned thresholds are retained as diagnostics rather than retrospectively reapplied to generate alarms. A validation audit found only rare numerical boundary disagreements between returned alarm indices and retrospective score–threshold comparisons; these did not alter the authoritative SPOT predictions.

Because F1 is structurally uninformative when an evaluation mixture contains no anomalies, the primary comparison uses the 320 conditions with positive target prevalence. All methods receive identical evaluation labels and condition sizes. Statistical comparisons use two-sided paired Wilcoxon

Table 6: Precision, recall, and F1 for the post-confirmatory five-method threshold comparison. Values are means over the 80 matched positive-prevalence seed–condition realizations per dataset–method combination. No comparator parameter was retuned after benchmark performance was observed.

Dataset	Method	Precision	Recall	F1
CICIDS2018	CAVET-AD	0.2205	0.0048	0.0086
	Fixed reference	0.2205	0.0048	0.0086
	Empirical $q = .99$	0.2199	0.0048	0.0086
	Robust MAD	0.1208	0.0091	0.0135
	Tukey-IQR	0.1067	0.0104	0.0147
CICIoT2023	CAVET-AD	0.7644	0.6697	0.7106
	Fixed reference	0.7818	0.2812	0.3939
	Empirical $q = .99$	0.7817	0.2812	0.3939
	Robust MAD	0.7960	0.1769	0.2750
	Tukey-IQR	0.7974	0.4473	0.5432
CreditCard2013	CAVET-AD	0.7930	0.8173	0.7998
	Fixed reference	0.8616	0.6112	0.7028
	Empirical $q = .99$	0.8616	0.6114	0.7029
	Robust MAD	0.7793	0.8293	0.7750
	Tukey-IQR	0.7392	0.8510	0.7561
NSL-KDD	CAVET-AD	0.7981	0.8366	0.8098
	Fixed reference	0.8757	0.7543	0.7995
	Empirical $q = .99$	0.8755	0.7543	0.7994
	Robust MAD	0.7742	0.8134	0.7634
	Tukey-IQR	0.7660	0.8174	0.7595

Table 8: Classification precision and recall for the exactly paired CAVET-AD, fixed-reference, and SPOT comparison. Values are means over 80 positive-prevalence seed–condition realizations per dataset. The corresponding F1 and paired statistical results are reported in Table 7.

	CAVET-AD		Fixed reference		SPOT	
Dataset	Precision	Recall	Precision	Recall	Precision	Recall
CICIDS2018	0.0217	0.0047	0.0217	0.0047	0.0225	0.0047
CICIoT2023	0.7573	0.6609	0.7788	0.2929	0.7862	0.2620
CreditCard2013	0.7886	0.8118	0.8583	0.5996	0.8552	0.6221
NSL-KDD	0.7953	0.8399	0.8755	0.7515	0.8385	0.7614

signed-rank tests, with Holm correction across the prespecified CAVET-versus-SPOT and CAVET-versus-fixed comparisons.

Table 7: Comparison with SPOT on exactly paired positive-prevalence conditions. Values are mean F1 $\pm$ standard deviation over 80 matched seed–prevalence conditions per dataset. Δ denotes CAVET-AD minus SPOT. p_{H} is the Holm-adjusted two-sided paired Wilcoxon p-value for CAVET-AD versus SPOT. Classification precision and recall are reported in Table 8.

Dataset	CAVET-AD	Fixed ref.	SPOT	Δ	p_{H}
CICIDS2018	0.0084 ± 0.0020	0.0084 ± 0.0020	0.0085 ± 0.0023	-0.0001	0.980
CICIoT2023	0.7018 ± 0.2031	0.4000 ± 0.1580	0.3600 ± 0.1035	$+0.3418$	< 0.001
CreditCard2013	0.7947 ± 0.1497	0.6929 ± 0.0965	0.7029 ± 0.1025	$+0.0918$	< 0.001
NSL-KDD	0.8099 ± 0.1308	0.7979 ± 0.1020	0.7795 ± 0.1342	$+0.0304$	< 0.001

Figure 1 shows the corresponding prevalence-resolved trajectories. CAVET-AD obtains a pooled descriptive mean F1 of 0.579, compared with 0.463 for SPOT and 0.475 for the fixed-reference rule. Relative to SPOT, the mean paired F1 differences are 0.342 on CICIoT2023, 0.092 on CreditCard2013, and 0.030 on NSL-KDD. These differences remain significant after Holm correction ($p_{\mathrm{H}} < 0.001$, $p_{\mathrm{H}} < 0.001$, and $p_{\mathrm{H}} < 0.001$, respectively).

On CICIDS2018, all methods remain near zero, and CAVET-AD is statistically indistinguishable from SPOT ($p_{\mathrm{H}} = 0.980$). Across the 320 positive-prevalence conditions, CAVET-AD exceeds SPOT in 240 conditions, ties in none, and is lower in 80. These findings support a conditional rather

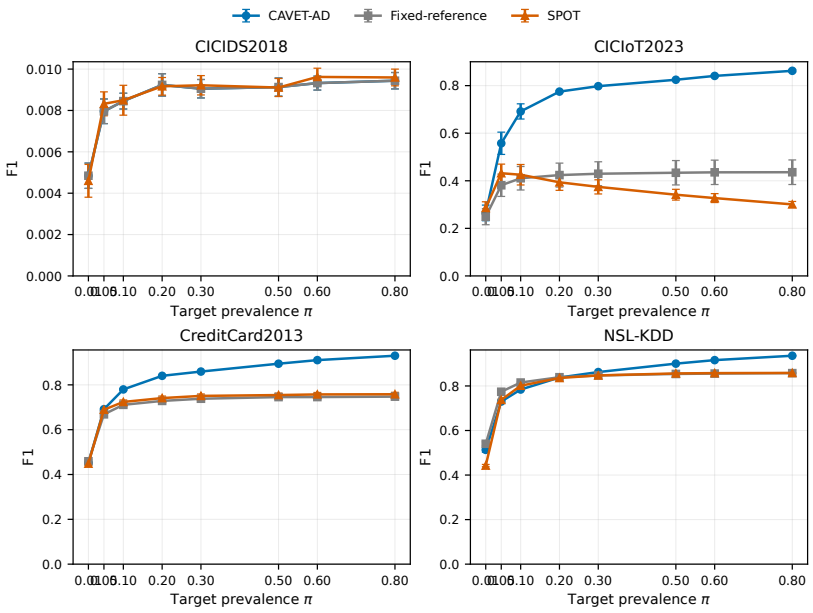


Figure 1: Mean F1 as a function of target anomaly prevalence for CAVET-AD, fixed-reference thresholding, and SPOT. Curves show means over ten exactly matched seeds; error bars denote one standard error. The zero-prevalence condition is excluded because F1 is not informative when no anomalies are present.

than universal advantage: adaptive tail calibration is beneficial when the detector-score tail retains informative prevalence-responsive structure, but it cannot repair an uninformative score geometry.

D.6 Comparison with Conformal Outlier Detection

We evaluate a direct conformal outlier-detection comparator following the multiple-testing formulation of Bates et al. (2023). The frozen oriented Isolation-Forest score is used as the nonconformity score. For an evaluation score s_i, its upper-tail conformal p-value is

$$p_i = \frac{1 + \sum_{j=1}^{n_{\mathrm{ref}}} \mathbb{I}\left(s_j^{\mathrm{ref}} \geq s_i\right)}{n_{\mathrm{ref}} + 1}.$$

The calibration sample contains only the frozen nominal reference scores. Within each deployment batch, Benjamini–Hochberg is applied at the prespecified false-discovery-rate level $q = 0.05$. The procedure uses neither deployment labels nor true prevalence. Its formal calibration interpretation depends on the corresponding exchangeability assumptions. This comparator addresses reference-calibrated conformal outlier testing with batch-level false-discovery-rate control; it is not a deployment-prevalence estimator and does not estimate the anomaly mixture proportion.

The experiment reuses the frozen detector scores, group-disjoint partitions, seeds, and controlled-prevalence mixtures from the post-confirmatory threshold-comparison study; no detector is trained or refitted. The analysis excludes the $\pi = 0$ all-normal condition because F1 is not informative without anomalies. This produces 80 exactly paired seed–prevalence conditions per dataset and 320 conditions overall. Dataset-level comparisons use two-sided paired Wilcoxon signed-rank tests, with Holm correction across the four prespecified hypotheses.

Across all 320 positive-prevalence conditions, the descriptive mean F1 is 0.5822 for CAVET-AD and 0.3809 for Conformal–BH, yielding a mean paired difference of $+0.2013$. CAVET-AD has higher F1 in 299 matched conditions, while Conformal–BH has higher F1 in 21; there are no exact ties. The largest mean difference occurs on CICIoT2023 ($+0.4334$), followed by CreditCard2013 ($+0.2677$), NSL-KDD ($+0.0956$), and CICIDS2018 ($+0.0086$).

Table 9: Comparison with Conformal–BH over the positive-prevalence protocol. Precision, recall, and F1 are means over 80 exactly matched seed–prevalence conditions per dataset. ΔF_1 denotes CAVET-AD minus Conformal–BH, and p_{H} is the Holm-adjusted two-sided paired Wilcoxon p-value.

	CAVET-AD			Conformal–BH				
Dataset	Precision	Recall	F1	Precision	Recall	F1	ΔF_1	p_{H}
CICIDS2018	0.2205	0.0048	0.0086	0.0000	0.0000	0.0000	+0.0086	< 0.001
CICIoT2023	0.7644	0.6697	0.7106	0.6923	0.2097	0.2772	+0.4334	< 0.001
CreditCard2013	0.7930	0.8173	0.7998	0.7153	0.4449	0.5321	+0.2677	< 0.001
NSL-KDD	0.7981	0.8366	0.8098	0.8228	0.6336	0.7142	+0.0956	< 0.001

Conformal–BH is especially conservative on CICIDS2018, where 85 of the 90 complete-grid conditions produce no alerts. This is not a score-computation failure: the conformal p-values do not cross the batch-adjusted rejection boundary. The comparison therefore distinguishes operating objectives rather than establishing universal method dominance. Conformal–BH controls a prespecified multiple-testing criterion under its assumptions, whereas CAVET-AD changes the operating budget in response to observable excess tail mass.

D.7 Component Ablation

The post-confirmatory ablation study compares three exactly paired operating rules: full CAVET-AD, a no-adaptation variant with the frozen budget $b = b_0 = 0.01$, and a no-EVT-refinement variant that retains the adaptive budget but uses empirical reference-survivor inversion. The same fitted detector, partitions, and evaluation realizations are retained across all three arms; no detector is refitted for the ablation.

The ablation shows that removing adaptive budgeting causes the dominant performance change on CICIoT2023 and CreditCard2013 and a smaller aggregate change on NSL-KDD, while CICIDS2018 remains essentially unchanged. In contrast, removing EVT refinement while retaining the adaptive budget produces only small aggregate F1 changes. Thus, the principal empirical contribution comes from adaptive budgeting, whereas EVT refinement primarily affects the reference-survivor model used for threshold inversion.

D.8 Complete Prevalence-Resolved Confirmatory Inference

Table 12 reports the complete prevalence-resolved paired inference from the frozen confirmatory experiment. For every dataset and prespecified nonzero prevalence condition, the table reports the mean paired effect $\Delta F_1 = F_{1,\mathrm{CAVET}} - F_{1,\mathrm{fixed}}$, its two-sided seed-level Student-t 95% confidence interval, the exact paired sign-flip p-value adjusted by Holm correction across the eight nonzero-prevalence hypotheses within that dataset, and the resulting inferential classification. All entries are computed from the same ten frozen confirmatory seeds. The $\pi = 0$ all-normal condition is retained elsewhere as an operating sanity check but is excluded from this effectiveness table and from the multiplicity-adjusted prevalence-resolved interpretation.

The precision–recall decomposition clarifies that the primary gains are driven mainly by increased recall. On CICIoT2023, CAVET-AD increases recall from 0.293 to 0.661 with only a modest precision reduction. On CreditCard2013, recall increases from 0.600 to 0.812, accompanied by a larger precision reduction. NSL-KDD shows a smaller recall increase, while CICIDS2018 remains unchanged. Thus, the F1 improvements do not arise from a uniform improvement in both component metrics; they represent dataset-dependent precision–recall tradeoffs induced by the adaptive operating point.

The prevalence-resolved inference yields four distinct regimes. CICIoT2023 is CAVET-favorable at all eight nonzero prevalence conditions. CreditCard2013 contains one directional but non-Holm-significant condition at $\pi = 0.01$ and seven CAVET-favorable conditions beginning at $\pi = 0.05$. NSL-KDD is fixed-favorable at $\pi = 0.01$, 0.05, and 0.10, unresolved at $\pi = 0.20$, and CAVET-favorable

Table 10: Mean F1 in the post-confirmatory component-ablation study. All three arms are evaluated on the same dataset–seed–prevalence realizations.

Dataset	Full CAVET	No adaptation	No EVT refinement
CICIDS2018	0.008570	0.008570	0.008568
CICIoT2023	0.710605	0.393912	0.710600
CreditCard2013	0.799844	0.702757	0.799853
NSL-KDD	0.809822	0.799528	0.809698

Table 11: Classification precision, recall, and F1 for the frozen primary Isolation Forest experiment. Values are means over the same $10 \times 8 = 80$ positive-prevalence seed–condition realizations per dataset. The $\pi = 0$ all-normal condition is excluded because positive-class precision and recall are not informative there.

	CAVET-AD			Fixed reference		
Dataset	Precision	Recall	F1	Precision	Recall	F1
CICIDS2018	0.0217	0.0047	0.0084	0.0217	0.0047	0.0084
CICIoT2023	0.7573	0.6609	0.7018	0.7788	0.2929	0.4000
CreditCard2013	0.7886	0.8118	0.7947	0.8583	0.5996	0.6929
NSL-KDD	0.7953	0.8399	0.8099	0.8755	0.7515	0.7979

from $\pi = 0.30$ onward. CICIDS2018 remains unresolved across all eight nonzero conditions because the paired F1 difference is identically zero.

D.9 Multiplicity-Adjusted Direction of the Effects

Figure 3 provides a compact view of the multiplicity-adjusted prevalence-specific conclusions.

Three distinct patterns are visible. CICIoT2023 is CAVET-favorable across all eight tested nonzero prevalence conditions. CreditCard2013 has a slightly negative directional effect at $\pi = 0.01$ that does not remain significant after Holm correction and is CAVET-favorable at the remaining seven conditions. NSL-KDD contains three fixed-favorable low-prevalence conditions, one unresolved transition condition, and four CAVET-favorable higher-prevalence conditions. CICIDS2018 remains unresolved across the complete sweep because the paired effect is exactly zero. Complete numerical results are reported in Table 12.

These patterns reinforce the interpretation that CAVET-AD is a regime-dependent operating rule rather than a universally dominating threshold.

D.10 Retrospective Tail-Identifiability Diagnostic

Observable tail response and adaptation gain. Figure 4 provides the prevalence-resolved diagnostic underlying the mechanism analysis in Section 5. It compares the observable excess-tail response r_{tail} with the F1 gain attributable to adaptive budgeting, $\Delta F_{1,\mathrm{adapt}}$.

As a complementary retrospective diagnostic, we also examine class-conditional tail identifiability using evaluation labels after completion of the frozen experiment. This analysis is *retrospective*: evaluation labels are used only after completion of the frozen confirmatory experiment to characterize how frequently nominal and anomalous observations exceed the frozen reference EVT threshold u. The diagnostic is not supplied to the deployed CAVET-AD rule and was not used for parameter selection.

For each dataset, let

$$P_N = P(Z > u \mid Y = 0), \qquad P_A = P(Z > u \mid Y = 1),$$

denote the empirical upper-tail occupancy of nominal and anomalous evaluation observations, respectively. We summarize their separation by

$$\Delta_{\mathrm{tail}} = P_A - P_N, \qquad R_{\mathrm{tail}} = \frac{P_A}{P_N}.$$

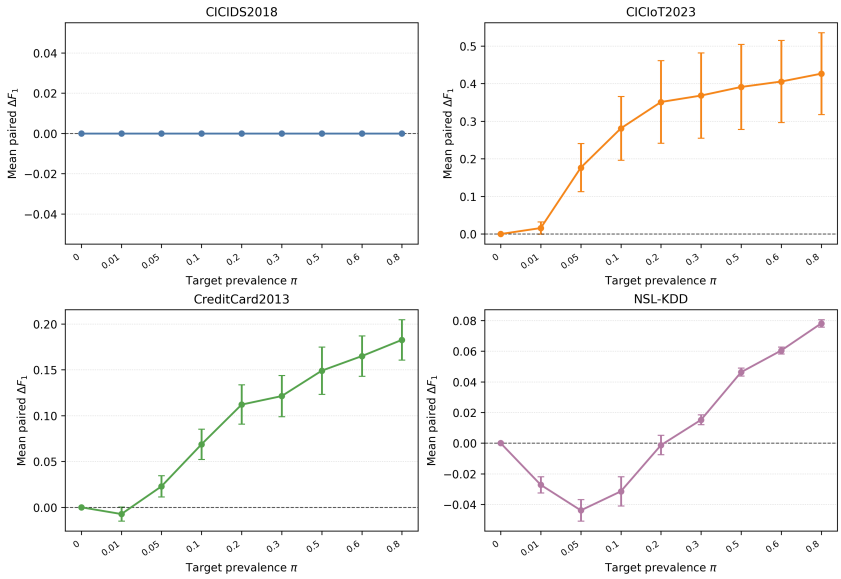


Figure 2: Prevalence-resolved paired ΔF_1 across ten frozen seeds. Positive values favor CAVET-AD. The $\pi = 0$ condition is an all-normal sanity check. Two-sided Student-t intervals and Holm-adjusted results are in Table 12.

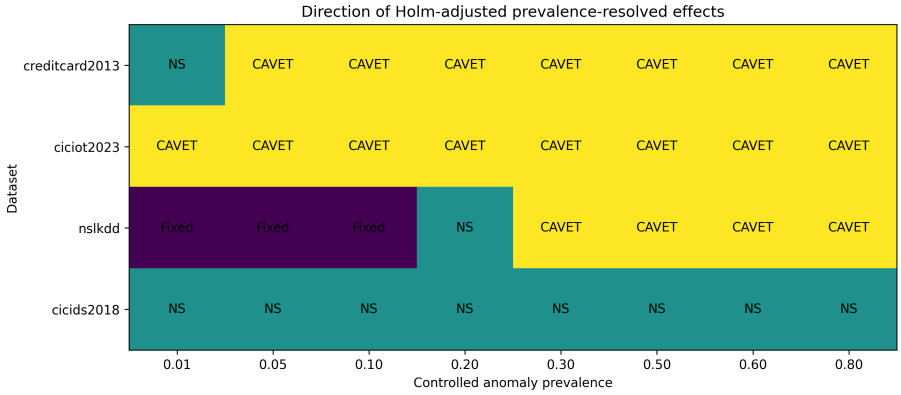


Figure 3: Direction of Holm-adjusted prevalence-resolved paired effects for the nonzero controlled prevalence conditions. "CAVET" denotes a statistically supported CAVET-AD advantage, "Fixed" denotes a statistically supported fixed-reference-tail advantage, and "NS" denotes a condition that remains statistically unresolved after multiplicity correction.

Table 13 reports the frozen diagnostic together with the corresponding aggregate confirmatory effect.

CreditCard2013, CICIoT2023, and NSL-KDD exhibit strong positive tail separation. Their anomalous upper-tail occupancies are approximately 0.850, 0.754, and 0.841, respectively, compared with nominal occupancies near 0.05. The resulting tail ratios are approximately 17.02, 15.21, and 16.83.

CICIDS2018 exhibits the opposite configuration. Its nominal upper-tail occupancy is approximately 0.050, whereas its anomalous upper-tail occupancy is only approximately 0.0095, yielding

$$\Delta_{\text{tail}} \approx -0.0405, \qquad R_{\text{tail}} \approx 0.189.$$

Descriptively, the three datasets with positive tail identifiability also exhibit positive aggregate paired F1 effects, whereas CICIDS2018 exhibits negative tail identifiability together with an exact zero aggregate paired effect. This correspondence is consistent with the interpretation that CAVET-AD benefits when anomalous observations produce a usable positive increase in reference-tail occupancy.

Table 12: Complete prevalence-resolved paired confirmatory inference. $\Delta F_1 = F_{1,\mathrm{CAVET}} - F_{1,\mathrm{fixed}}$; positive values favor CAVET-AD and negative values favor the fixed reference-tail rule. CI denotes the two-sided seed-level Student-t 95% confidence interval, and Holm p denotes the exact paired sign-flip p-value adjusted across the eight displayed conditions within the corresponding dataset. Only the eight prespecified nonzero prevalence conditions are shown; $\pi = 0$ is an all-normal sanity condition and is not an effectiveness comparison.

Dataset	π	Mean ΔF_1	95% CI	Holm p	Sig.	Interpretation
CreditCard2013						
	0.01	-0.0073	[-0.0150, 0.0003]	0.0586	No	Directional, not Holm-significant
	0.05	0.0229	[0.0114, 0.0343]	0.0156	Yes	CAVET favorable
	0.10	0.0687	[0.0522, 0.0852]	0.0156	Yes	CAVET favorable
	0.20	0.1120	[0.0906, 0.1334]	0.0156	Yes	CAVET favorable
	0.30	0.1213	[0.0988, 0.1437]	0.0156	Yes	CAVET favorable
	0.50	0.1490	[0.1232, 0.1747]	0.0156	Yes	CAVET favorable
	0.60	0.1649	[0.1428, 0.1870]	0.0156	Yes	CAVET favorable
	0.80	0.1825	[0.1606, 0.2045]	0.0156	Yes	CAVET favorable
CICIDS2018						
	0.01	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
	0.05	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
	0.10	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
	0.20	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
	0.30	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
	0.50	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
	0.60	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
	0.80	0.0000	[0.0000,0.0000]	1.0000	No	Unresolved
CICIoT2023						
	0.01	0.0158	[-0.0001,0.0318]	0.0313	Yes	CAVET favorable
	0.05	0.1765	[0.1127,0.2404]	0.0156	Yes	CAVET favorable
	0.10	0.2810	[0.1961,0.3658]	0.0156	Yes	CAVET favorable
	0.20	0.3510	[0.2410,0.4609]	0.0156	Yes	CAVET favorable
	0.30	0.3681	[0.2549,0.4812]	0.0156	Yes	CAVET favorable
	0.50	0.3909	[0.2780,0.5038]	0.0156	Yes	CAVET favorable
	0.60	0.4053	[0.2961,0.5146]	0.0156	Yes	CAVET favorable
	0.80	0.4263	[0.3174,0.5351]	0.0156	Yes	CAVET favorable
NSL-KDD						
	0.01	-0.0272	[-0.0325,-0.0219]	0.0156	Yes	Fixed favorable
	0.05	-0.0439	[-0.0508,-0.0369]	0.0156	Yes	Fixed favorable
	0.10	-0.0315	[-0.0410,-0.0219]	0.0156	Yes	Fixed favorable
	0.20	-0.0013	[-0.0076, 0.0050]	0.6387	No	Unresolved
	0.30	0.0152	[0.0120, 0.0184]	0.0156	Yes	CAVET favorable
	0.50	0.0463	[0.0437, 0.0489]	0.0156	Yes	CAVET favorable
	0.60	0.0604	[0.0581, 0.0627]	0.0156	Yes	CAVET favorable
	0.80	0.0781	[0.0757, 0.0805]	0.0156	Yes	CAVET favorable

Table 13: Post-confirmatory tail-identifiability diagnostic. P_N and P_A denote the empirical probabilities that nominal and anomalous evaluation observations, respectively, exceed the frozen EVT reference threshold u. We define $\Delta_{\mathrm{tail}} = P_A - P_N$ and $R_{\mathrm{tail}} = P_A/P_N$. AUROC and AP characterize the frozen detector ranking, while ΔF_1 and Holm p summarize the frozen confirmatory operating-point comparison. Evaluation labels are used only retrospectively for this diagnostic and were not supplied to CAVET-AD or used for parameter selection. The reported ΔF_1 values reproduce the complete nine-condition dataset-level summaries from Table 1.

Dataset	P_N	P_A	Δ_{tail}	R_{tail}	AUROC	AP	ΔF_1	Holm p	Tail ID
CreditCard2013	0.0499	0.8496	0.7997	17.023	0.9475	0.1643	0.0904	0.0078	Positive
CICIDS2018	0.0500	0.0095	-0.0405	0.189	0.5162	0.1817	0.0000	1.0000	Negative
CICIoT2023	0.0496	0.7536	0.7041	15.208	0.9544	0.9991	0.2683	0.0078	Positive
NSL-KDD	0.0499	0.8406	0.7907	16.833	0.9687	0.9799	0.0107	0.0078	Positive

However, the diagnostic is retrospective and should not be interpreted as a causal explanation or as a quantitative predictor of the magnitude of F1 improvement. In particular, CICIoT2023 produces the largest confirmatory gain even though its Δ_{tail} is smaller than those of CreditCard2013 and NSL-KDD. The diagnostic therefore provides a mechanistic characterization of the observed operating regimes rather than a deployment criterion.

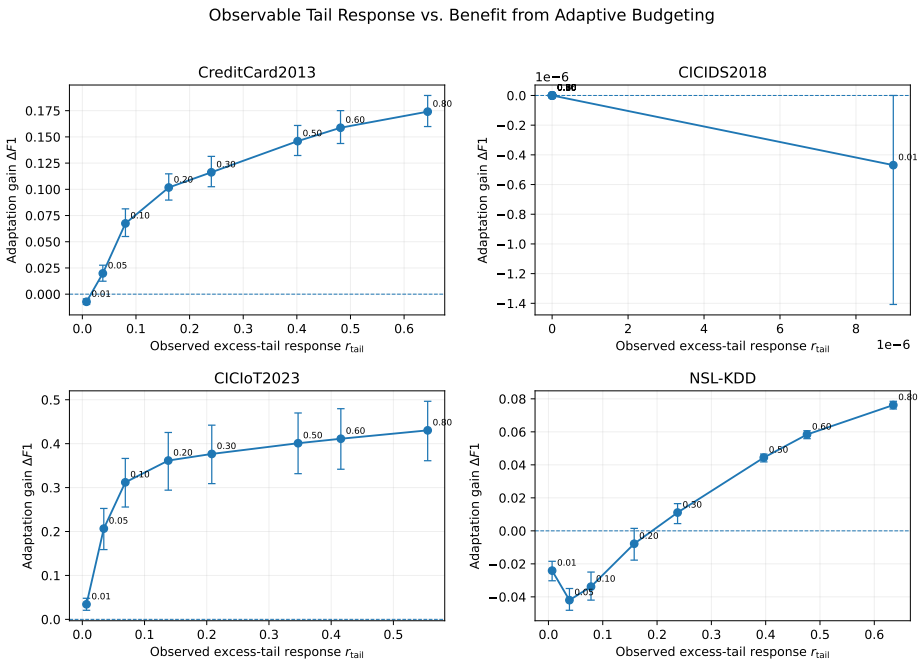


Figure 4: Observable excess-tail response versus the F1 benefit attributable to adaptive budgeting. Each panel shows prevalence-resolved mean behavior for one dataset; error bars summarize variation across the ten frozen confirmatory seeds. CreditCard2013 and CICIoT2023 exhibit strong positive alignment, while NSL-KDD shows a low-response negative regime followed by a positive adaptive crossover. CICIDS2018 supplies effectively no usable positive tail-response variation and correspondingly no meaningful adaptation gain.

D.11 Detailed Held-Out Evaluation of the Applicability Signal

The preceding analysis establishes a contemporaneous association between the observable excess-tail response r_{tail} and the benefit of adaptive budgeting. We next evaluate whether this signal can support a decision for a dataset that was not used to fit the decision cutoff.

We conducted a leave-one-dataset-out policy evaluation across the four confirmatory benchmarks. In each fold, an r_{tail} cutoff was selected using only the other three development datasets and was then frozen before evaluation on the held-out dataset. The policy applies CAVET-AD when r_{tail} exceeds the frozen cutoff and otherwise retains the fixed reference-tail rule. Labeled F_1 outcomes were used offline to select the cutoff on the development datasets, but neither labels nor true prevalence from the held-out dataset were used to select the cutoff or make its policy decisions. The $\pi = 0$ all-normal sanity condition was excluded, leaving 320 held-out dataset–seed–prevalence decisions and 40 independent dataset–seed inference units. All four folds passed the leakage audit.

Table 14: Leave-one-dataset-out evaluation of the r_{tail} applicability policy. Adaptive denotes always applying CAVET-AD. Gate selects between the fixed and adaptive rules using a cutoff fitted on the other three datasets. Each dataset entry is averaged over ten seeds and eight nonzero prevalence conditions.

Held-out dataset	Fixed F_1	Adaptive F_1	Gate F_1	Adapt rate
CICIDS2018	0.0084	0.0084	0.0084	0.000
CICIoT2023	0.4000	0.7018	0.6778	0.750
CreditCard2013	0.6929	0.7947	0.7956	0.875
NSL-KDD	0.7979	0.8099	0.8101	0.988
Overall	0.4748	0.5787	0.5730	0.653

Across the 40 dataset–seed units, the held-out gate improves mean F_1 over the fixed rule by 0.0982 (two-sided Wilcoxon $p = 2.14 \times 10^{-7}$). Its mean F_1 is 0.0057 below the always-adaptive policy, but this difference is not statistically significant ($p = 0.131$). The gate selects the oracle action in 83.1% of the 320 held-out conditions.

The most direct no-adaptation result occurs on CICIDS2018: the frozen cutoff retains the fixed rule in every held-out condition and exactly preserves its performance. On CICIoT2023, the gate substantially outperforms the fixed rule but remains 0.0240 below always adapting, showing that the cross-dataset cutoff does not recover every beneficial adaptation decision. On CreditCard2013 and NSL-KDD, gate performance is approximately equal to the always-adaptive result.

This experiment is a cross-fitted evaluation using frozen confirmatory outputs, not a live temporal deployment. It therefore supports cross-dataset transfer of the applicability signal under the evaluated pure prevalence-shift protocol, but does not establish robustness to future temporal, covariate, or concept drift.

D.12 Threats to Validity

We follow standard empirical methodology and evaluate threats to: (i) internal validity (base detectors, EVT fitting), (ii) external validity (dataset diversity), (iii) construct validity (score alignment), and (iv) statistical conclusion validity (seed variance, significance tests).

- **Internal validity:** EVT tail fitting can be unstable when the frozen reference partition contains too few upper-tail exceedances. This risk is evaluated through the post-confirmatory EVT-fitting stability diagnostics reported in the paper.
- **External validity:** CAVET-AD is evaluated on tabular cybersecurity and financial datasets; vision and NLP domains remain future work.
- **Construct validity:** The reference-CDF transformation maps detector scores into a common reference-tail coordinate system. Its purpose is operating-point calibration rather than modification of the underlying score ranking.
- **Statistical conclusion validity:** The confirmatory analysis uses ten prespecified seeds, paired comparisons, two-sided paired-seed Student-t confidence intervals, and exact seed-level sign-flip inference. Holm correction is applied across the four aggregate dataset hypotheses and, for prevalence-resolved inference, separately within each dataset across its eight prespecified nonzero-prevalence hypotheses. For m nonzero paired differences, the smallest attainable two-sided sign-flip p-value is 2^{1-m}. Thus, at $m = 10$, the raw floor is 0.001953125, while the best-case Holm-adjusted floors are 0.0078125 for the four-test aggregate family and 0.015625 for an eight-test within-dataset family; ties make the inference still coarser. No seeds were added after inspection of the frozen confirmatory outcomes. Finer confirmatory resolution requires a separately prespecified larger-seed replication. The seed, rather than the individual evaluation observation, is the interval-estimation and inferential unit.

Additional scope limitations. The empirical scope remains limited to four tabular benchmarks. The full-protocol detector evidence covers Isolation Forest and ECOD, with LOF examined only in the supplementary transfer analysis; deep detectors, representation-learning methods, and non-tabular modalities remain unevaluated. Natural-prevalence evidence is limited to CreditCard2013, and the covariate-shift study uses one controlled feature-translation experiment rather than natural temporal, concept, or combined prevalence–covariate shift.

CAVET-AD also assumes a representative nominal reference sample; robustness to reference contamination remains untested. Its linear budget function is heuristic: it performs best among the prespecified schedules examined here but has no decision-theoretic optimality guarantee. Comparator coverage includes one reference-calibrated Conformal–BH procedure, but not a direct empirical prevalence estimator. In addition, the applicability cutoff is selected offline using labeled outcomes from the development datasets. Leave-one-dataset-out evaluation prevents leakage from the held-out dataset, but does not constitute prospective temporal validation.

The experiments also identify boundaries that limit any universal improvement claim: the CreditCard2013 effect is slightly negative at $\pi = 0.01$, including under ECOD; its natural-prevalence

improvement is not significant after multiplicity correction; and the benefit varies across detectors and datasets.

D.13 Independent Nominal Calibration in Normalized Z-Space

CAVET-AD maps oriented anomaly scores into a common reference-tail space,

$$Z = -\log\bigl(1 - \widehat{F}_{\mathrm{ref}}(s)\bigr),$$

where $\widehat{F}_{\mathrm{ref}}$ is estimated exclusively from the frozen reference-score partition.

Applying this empirical CDF transformation back to the same reference sample produces near-ideal rank-based agreement by construction and is therefore useful only as an implementation sanity check. To obtain a non-circular calibration diagnostic, we instead apply the frozen reference CDF to *independent nominal observations from the disjoint evaluation partition*.

The detector and preprocessing pipeline remain fitted only on the detector-training partition, and the empirical CDF remains fitted only on the frozen reference partition. Evaluation labels are used solely to identify nominal observations for this retrospective diagnostic; they are not supplied to the detector, CAVET-AD operating rule, EVT fit, threshold selection, or any parameter-selection procedure.

Under an adequate reference-space calibration, these independent nominal transformed scores should exhibit approximate agreement with the $\mathrm{Exp}(1)$ reference distribution. Exact exponential behavior is not assumed, particularly in the extreme upper tail.

Independent-Nominal QQ Diagnostic

Interpretation Figure 5 shows that the independently evaluated nominal Z distributions track the theoretical $\mathrm{Exp}(1)$ reference closely over most of their range. CreditCard2013, CICIDS2018, and CICIoT2023 show particularly close agreement, whereas NSL-KDD exhibits a modest departure in the extreme upper tail. Because the observations used for this diagnostic are disjoint from those used to construct $\widehat{F}_{\mathrm{ref}}$, the agreement is not a rank-calibration identity induced by re-evaluating the reference sample itself.

The diagnostic therefore supports the intended reference-space normalization while also showing that exact exponential behavior should not be assumed in every dataset or throughout the extreme tail. The remaining upper-tail departures motivate the EVT refinement used by CAVET-AD for tail-sensitive threshold modeling.

D.14 Independent-Nominal Calibration Goodness of Fit

To quantify the independent calibration diagnostic in Section D.13, we compute a descriptive Kolmogorov–Smirnov (KS)-type discrepancy between the empirical CDF of the independent nominal transformed scores and the theoretical $\mathrm{Exp}(1)$ reference:

$$D = \sup_{z} \left| \widehat{F}_{Z,\text{eval-nom}}(z) - \left(1 - e^{-z}\right) \right|,$$

where $\widehat{F}_{Z,\text{eval-nom}}$ is computed exclusively from nominal observations in the disjoint evaluation partition after applying the frozen reference ECDF.

The diagnostic uses the same prespecified confirmatory seed ($s = 7$) and the same independent nominal Z samples used for Figure 5. No controlled-prevalence mixture, anomaly evaluation observation, performance metric, or parameter-selection procedure enters this calculation. Because the reference transformation is itself empirically estimated, D is interpreted descriptively rather than as a formal one-sample KS hypothesis test; no KS p-value is used.

Interpretation All four independent nominal distributions remain close to the $\mathrm{Exp}(1)$ reference in their first two moments and exhibit small descriptive KS-type discrepancies. CICIoT2023 yields the smallest discrepancy ($D = 0.0044$), followed by NSL-KDD ($D = 0.0063$) and CreditCard2013 ($D = 0.0064$). CICIDS2018 exhibits the largest discrepancy ($D = 0.0246$), although its transformed-score mean and standard deviation remain very close to the theoretical $\mathrm{Exp}(1)$ values of one.

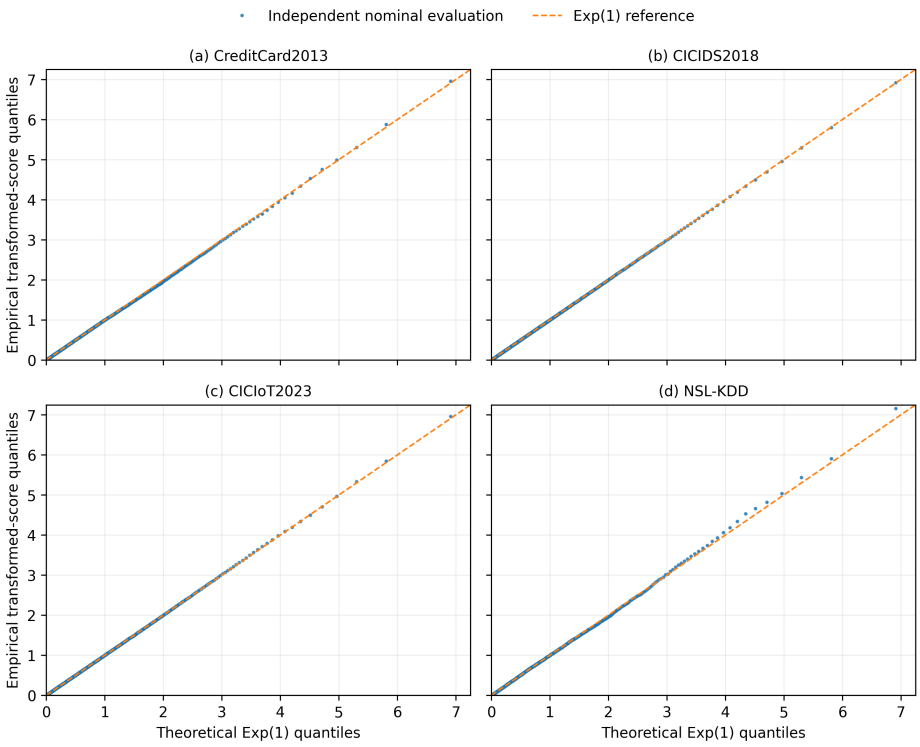


Figure 5: Independent nominal calibration diagnostic. Quantile–quantile comparison of transformed anomaly scores against the theoretical $\mathrm{Exp}(1)$ reference distribution for (a) CreditCard2013, (b) CICIDS2018, (c) CICIoT2023, and (d) NSL-KDD. The reference empirical CDF is constructed exclusively from the frozen reference partition, whereas the empirical quantiles shown here are computed from nominal observations in the disjoint evaluation partition. The close agreement across datasets supports the intended reference-space calibration, while the modest upper-tail deviation for NSL-KDD illustrates residual dataset-specific departure from the idealized reference distribution.

Table 15: Independent-nominal calibration goodness-of-fit against the $\mathrm{Exp}(1)$ reference. The reference ECDF is fitted only on the frozen reference partition, whereas D is evaluated on nominal observations from the disjoint evaluation partition using prespecified confirmatory seed 7. Lower values indicate closer distributional agreement. The statistic is descriptive and is not used for parameter selection.

Dataset	n_{nom}	D	$\mathrm{E}[Z]$	$\mathrm{SD}[Z]$
CreditCard2013	170,589	0.0064	0.9933	0.9917
CICIDS2018	7,147,231	0.0246	0.9999	0.9985
CICIoT2023	630,825	0.0044	1.0029	1.0033
NSL-KDD	46,234	0.0063	1.0003	1.0089

The QQ diagnostic in Figure 5 provides complementary information not captured by the scalar D: NSL-KDD shows a modest extreme-upper-tail departure despite a small overall KS-type discrepancy. Taken together, the quantitative and graphical diagnostics support close reference-space agreement over most of the nominal distribution while showing that exact exponential behavior should not be assumed uniformly in the extreme tail.

Table 16: Statistics and natural anomaly prevalence of the canonical processed datasets used in the experiments.

Dataset	Domain	#Samples	#Features	Contamination	Role
Credit Card EU 2013	Financial fraud	284,807	30	0.1727%	Tests threshold calibration in an extreme rare-anomaly tabular setting.
CICIDS2018	Enterprise/cloud traffic	13,522,722	79	11.9109%	Evaluates threshold adaptation under large-scale network traffic and prevalence shift.
CICIoT2023	IoT smart-home attacks	45,019,243	39	97.6646%	Provides an extreme high-prevalence benchmark for evaluating adaptive threshold behavior.
NSL-KDD	Network intrusion	148,517	41	48.1177%	Evaluates adaptive threshold behavior in a moderate-to-high prevalence intrusion setting.

E Confirmatory Experimental Protocol

The final confirmatory experiment follows the frozen design described in Section 4. For every dataset and prespecified random seed, observations are assigned to mutually exclusive *row-index* detector-training, reference, and evaluation partitions.

Table 16 summarizes the canonical processed datasets used in the confirmatory evaluation.

The detector-training partition is used only to fit the base anomaly detector. The reference partition is used only after detector fitting to construct the frozen reference-score distribution, empirical reference CDF, and EVT-refined reference survivor. Controlled anomaly-prevalence mixtures are generated exclusively from the held-out evaluation pool.

The primary CAVET-AD configuration is fixed at

$$q = 0.95, \qquad b_0 = 0.01, \qquad \kappa = 0.20, \qquad b_{\max} = 0.10.$$

The ten prespecified confirmatory seeds are

$$S = \{7, 11, 13, 21, 23, 37, 51, 55, 73, 97\},$$

and the controlled prevalence conditions are

$$\Pi = \{0.00, 0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.60, 0.80\}.$$

This primary partitioning guarantees row-index separation. Because identical feature-content vectors may occur at different row indices in the canonical processed datasets, content-level separation was evaluated separately in a post-confirmatory group-disjoint robustness experiment.

For each dataset–seed–prevalence condition, CAVET-AD and the frozen fixed reference-tail operating rule are evaluated on the identical realized evaluation mixture. The primary paired effect is

$$\Delta F_1 = F_{1,\mathrm{CAVET}} - F_{1,\mathrm{fixed}}.$$

Post-confirmatory EVT-anchor sensitivity, adaptive-budget sensitivity, EVT-fitting stability, and tail-identifiability diagnostics are performed only after the primary configuration and confirmatory conclusions have been frozen. None of these analyses is used for retrospective parameter selection.

F Computational Complexity

This section analyzes the computational efficiency of CAVET-AD, covering theoretical complexity, empirical runtime, scaling behavior, and memory/streaming feasibility. Although CAVET-AD adds a

Table 17: Per-stage complexity of CAVET-AD.

Stage	Operation	Complexity	Notes
Base detector	score generation $s = f(x)$	model-specific	unchanged by CAVET-AD
Reference CDF	sort frozen reference scores	$O(N_{\mathrm{ref}} \log N_{\mathrm{ref}})$	performed once per reference partition
Tail transform	$Z = -\log(1 - \widehat{F}_{\mathrm{ref}}(s))$	$O(N)$	element-wise transformation
EVT tail fit	fit GPD to reference exceedances	$O(N_u)$	$N_u \ll N_{\mathrm{ref}}$
Tail response	compute r_{tail} from exceedances	$O(N)$	batch-level calibration statistic
Adaptive budget	$b = \min(b_{\max}, b_0 + \kappa r_{\mathrm{tail}})$	$O(1)$	after r_{tail} is available
Threshold inversion	$\tau^* = \bar{F}^{-1}_{\mathrm{ref,EVT}}(b)$	$O(1)$	EVT survivor inversion

Table 18: Measured CPU wall-clock latency of the complete post-hoc CAVET-AD batch decision on CICIDS2018 seed 7.

Batch size	Repetitions	Median (ms)	5th–95th pct. (ms)
1,000	300	0.044	0.044–0.051
10,000	200	0.845	0.807–1.061
100,000	100	16.290	15.946–17.487
859,950	30	209.089	205.726–214.341

calibration and EVT stage on top of a base detector, it maintains competitive cost relative to standard unsupervised pipelines.

F.1 Theoretical Complexity

Let N denote the number of incoming scores, N_{ref} the reference-partition size, and N_u the number of EVT exceedances. The main preprocessing cost introduced by CAVET-AD is construction of the sorted reference distribution. Score transformation and tail-response calculation are linear in the evaluated batch size, whereas adaptive-budget calculation and threshold inversion are constant-time once their required statistics are available.

Overall Complexity The resulting complexity is

$$T_{\mathrm{CAVET}} = O(T_{\mathrm{detector}} + N_{\mathrm{ref}} \log N_{\mathrm{ref}} + N + N_u)\,.$$

F.2 Measured Wall-Clock Runtime

We measured the complete post-hoc CAVET-AD stage using the frozen CICIDS2018 seed-7 score artifact, containing $N_{\mathrm{ref}} = 2{,}377{,}829$ reference scores and $859{,}950$ deployment scores. Measurements were obtained on an Intel Xeon Gold 6534 host with 32 logical CPU cores and approximately 503 GiB RAM, using Python 3.9.18, NumPy 2.0.2, and SciPy 1.13.1. The median one-time reference-calibration cost was 0.898 seconds over ten runs (5th–95th percentile: 0.882–0.947 seconds). This setup measurement includes reference-ECDF construction, reference-score transformation, GPD fitting, and preparation of the reference survivor for inversion.

Table 18 reports the measured decision latency across four deployment-batch sizes.

The per-batch measurement includes score-to-Z transformation, tail-response calculation, adaptive-budget calculation, reference-survivor inversion, and binary prediction. For the complete $859{,}950$-score deployment pool, the median latency was 209.089 milliseconds. Disk input/output and base-detector fitting and score generation are excluded because these operations precede the post-hoc CAVET-AD calibration stage and are detector-specific.

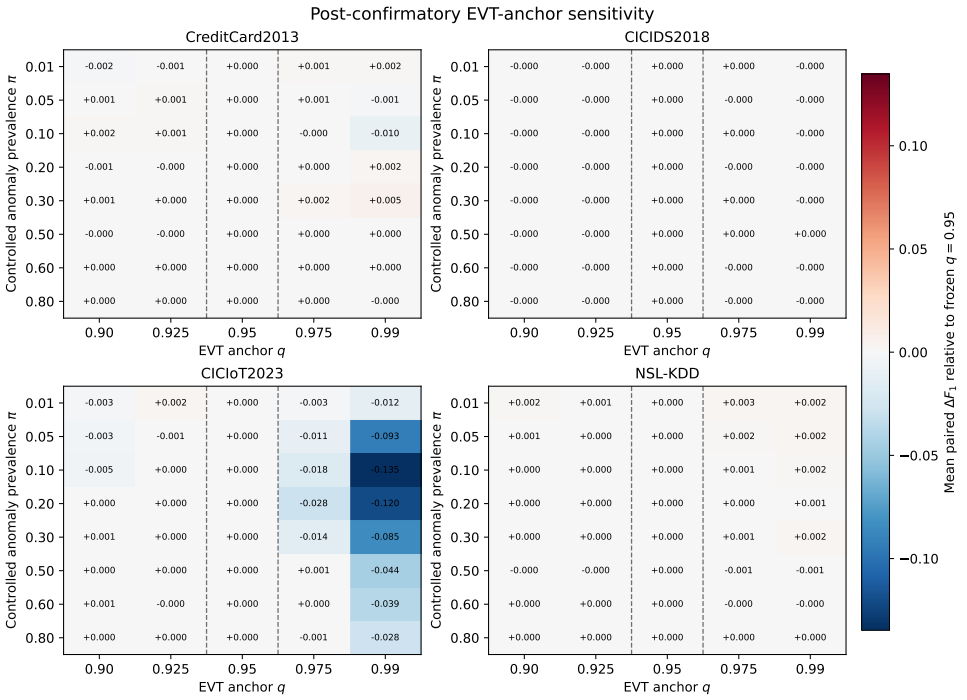


Figure 6: Post-confirmatory EVT-anchor sensitivity across the four confirmatory datasets. Each cell reports the mean paired F1 difference relative to the frozen primary anchor $q = 0.95$ for one controlled nonzero-prevalence condition. The dashed vertical boundaries identify the frozen $q = 0.95$ column. Values near zero indicate local robustness to anchor perturbation. CICIDS2018 is effectively invariant, while CreditCard2013 and NSL-KDD show only small changes. CICIoT2023 exhibits greater sensitivity at the extreme $q = 0.99$ setting, including a decrease of approximately 0.135 in F1 at $\pi = 0.10$. The perturbation results were not used to replace or re-select the frozen primary anchor.

G Post-Confirmatory Robustness Analyses

The analyses in this section were conducted only after the primary confirmatory configuration and conclusions had been frozen. They are therefore interpreted as robustness diagnostics rather than as a hyperparameter-selection procedure.

G.1 EVT-Anchor Sensitivity

The EVT anchor was evaluated over the prespecified local grid

$$q \in \{0.90, 0.925, 0.95, 0.975, 0.99\},$$

with $q = 0.95$ retained as the frozen primary value. For each dataset, anchor, and controlled prevalence, the same ten confirmatory seeds were used. Sensitivity is expressed as the paired difference

$$\Delta F_1(q) = F_1(q) - F_1(q = 0.95).$$

The $\pi = 0$ all-normal condition remains available as an operating sanity check but is excluded from the effectiveness visualization.

CICIDS2018 exhibits negligible anchor sensitivity across the complete tested range. CreditCard2013 and NSL-KDD are also locally stable around $q = 0.95$, with mean absolute F1 perturbations of approximately 0.0012 and 0.0007, respectively. CICIoT2023 is more sensitive: its mean absolute perturbation is approximately 0.0203, with the strongest departure occurring at $q = 0.99$ and $\pi = 0.10$.

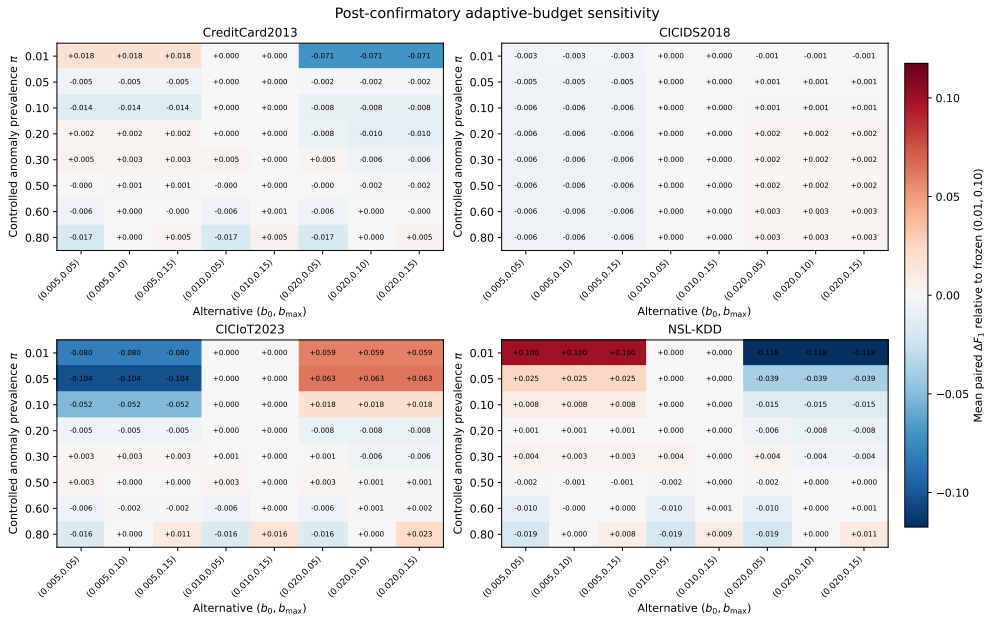


Figure 7: Post-confirmatory adaptive-budget sensitivity across the four confirmatory datasets. Each cell reports the mean paired F1 difference relative to the frozen primary configuration $(b_0, b_{\max}) = (0.01, 0.10)$ for one controlled nonzero-prevalence condition. Alternative configurations perturb the base and maximum tail budgets while keeping $q = 0.95$ and $\kappa = 0.20$ fixed. Values near zero indicate robustness to the corresponding budget perturbation. CICIDS2018 is comparatively stable, whereas CICIoT2023 and NSL-KDD exhibit larger localized departures. The perturbation results were not used to replace or re-select the frozen primary configuration.

These results support local robustness of the frozen anchor for three datasets while revealing an extreme-anchor sensitivity regime for CICIoT2023. The analysis is therefore interpreted as robustness characterization rather than as evidence for retrospective hyperparameter selection.

G.2 Adaptive-Budget Sensitivity

The adaptive-budget sensitivity analysis perturbs the base and maximum tail budgets around the frozen primary configuration

$$(b_0, b_{\max}) = (0.01, 0.10),$$

while holding the EVT anchor and adaptation coefficient fixed at $q = 0.95$ and $\kappa = 0.20$, respectively. The eight nonprimary configurations are

$$b_0 \in \{0.005, 0.010, 0.020\}, \qquad b_{\max} \in \{0.05, 0.10, 0.15\},$$

excluding the frozen primary pair $(0.01, 0.10)$.

For each alternative configuration, F1 was paired with the frozen primary result at the same dataset, confirmatory seed, and controlled prevalence. Sensitivity is therefore expressed as

$$\Delta F_1(b_0, b_{\max}) = F_1(b_0, b_{\max}) - F_1(0.01, 0.10).$$

The $\pi = 0$ all-normal condition was retained as a sanity condition but excluded from the effectiveness visualization.

Across the eight nonprimary configurations, CICIDS2018 exhibits the smallest overall sensitivity, with mean absolute paired F1 change approximately 0.0027 and maximum absolute change approximately 0.0109. CreditCard2013 has mean absolute change approximately 0.0089, with a largest localized departure of approximately 0.0824.

CICIoT2023 is the most sensitive dataset on average, with mean absolute paired F1 change approximately 0.0250 and maximum absolute change approximately 0.2281. NSL-KDD also exhibits

Table 19: Pooled performance of the prespecified budget schedules across four datasets and ten paired seeds. The linear schedule is the frozen primary rule. Results summarize dataset–seed units over the nine prevalence conditions.

Schedule	Mean F1	Median F1	F1 SD	Mean budget
Linear	**0.5822**	**0.7664**	0.3379	0.0432
Saturating ($\lambda = 3$)	0.5725	0.7493	0.3314	0.0535
Concave ($\gamma = 0.5$)	0.5700	0.7438	0.3294	0.0524
Convex ($\gamma = 2$)	0.5679	0.7526	0.3345	0.0360
Fixed	0.4762	0.6185	0.3174	0.0100

appreciable localized sensitivity, with mean absolute change approximately 0.0173 and maximum absolute change approximately 0.1496.

At the configuration level, the perturbations involving the smallest base budget or the largest base budget produce larger average departures than the configurations immediately surrounding the frozen primary setting. Nevertheless, the median paired changes remain close to zero for several configurations, indicating that the sensitivity is concentrated in particular dataset–prevalence regimes rather than being uniform across the factorial grid.

Accordingly, this experiment is interpreted as a robustness characterization of the frozen operating rule rather than as a hyperparameter search. No alternative $(b_0, b_{\max})$ pair was selected, and the primary configuration $(0.01, 0.10)$ remains unchanged.

G.3 Prespecified Nonlinear Budget Schedules

We additionally tested whether the linear mapping from excess-tail response to alert budget was unnecessarily restrictive. Using the frozen detector scores, group-disjoint partitions, EVT reference models, ten paired seeds, and nine prevalence conditions, we compared the primary linear schedule against three prespecified alternatives: a concave power rule ($\gamma = 0.5$), a convex power rule ($\gamma = 2$), and a saturating exponential rule ($\lambda = 3$). A fixed-budget rule was retained as a reference. Schedule parameters were specified before comparative evaluation; no detector was retrained, no partition was changed, and labels were not used to select a schedule. Reconstruction of all 360 frozen validation conditions matched the archived tail responses to numerical precision, with maximum absolute error 9.72×10^{-17}.

Table 19 shows that the frozen linear schedule has the highest pooled mean and median F1. Exact two-sided paired sign-flip tests were conducted within each dataset and corrected jointly over the 12 prespecified dataset–nonlinear-schedule comparisons using Holm's procedure. No nonlinear schedule differed significantly from linear on CICIDS2018. On CICIoT2023, the convex rule was significantly worse than linear (Holm $p = 0.0234$), while the concave and saturating rules were not significant. All three nonlinear rules were significantly worse on CreditCard2013 (Holm $p = 0.0234$ each). On NSL-KDD, the convex rule improved mean F1 by 0.0118, whereas the concave and saturating rules reduced it; all three comparisons were Holm-significant ($p = 0.0234$).

Thus, nonlinear scheduling can help a particular detector–dataset regime, but no tested nonlinear form improves consistently across datasets. The linear ramp is retained as the transparent default because it has the strongest pooled performance and avoids an additional shape parameter. This experiment does not establish universal optimality of the linear form or rule out a label-free learned schedule.

G.4 EVT-Fitting Stability

The reference-tail EVT fit was examined across the ten independently constructed confirmatory reference partitions for each dataset. This post-confirmatory diagnostic evaluates whether the primary findings could plausibly be explained by instability of the fitted reference-tail model. No detector, partition, confirmatory result, or operating parameter was re-estimated or selected from this analysis.

Table 20 summarizes the fitted EVT quantities across the ten frozen confirmatory seeds. The anchor threshold u, empirical reference exceedance probability p_u, GPD shape parameter ξ, and scale parameter σ are reported as mean $\pm$ standard deviation.

Table 20: Post-confirmatory stability of the frozen reference-tail EVT fits across the ten confirmatory seeds. Values are mean ± standard deviation. EVT fitting is performed only on the reference partition at the frozen anchor $q = 0.95$. All 40 dataset–seed fits used the same maximum-likelihood fitting mode. This analysis is diagnostic and was not used for parameter selection.

Dataset	Seeds	u	p_u	ξ	σ	Exceedances
CreditCard2013	10	2.9954 ± 0.0000	0.050015 ± 0.000000	-0.005358 ± 0.000000	1.003669 ± 0.000000	2844.0 ± 0.0
CICIDS2018	10	2.9957 ± 0.0001	0.049999 ± 0.000003	-0.000698 ± 0.000177	1.001331 ± 0.000302	119118.4 ± 6.9
CICIoT2023	10	2.9956 ± 0.0000	0.050001 ± 0.000000	-0.001963 ± 0.000052	1.001534 ± 0.000047	10514.0 ± 0.0
NSL-KDD	10	2.9948 ± 0.0006	0.050013 ± 0.000044	-0.014182 ± 0.001645	1.009587 ± 0.001502	770.7 ± 0.7

Table 21: Controlled covariate-shift stress test on CreditCard2013 at fixed prevalence $\pi = 0.10$ across ten paired seeds. Shift denotes translation magnitude along the leading training-PCA direction. KS measures displacement from the unshifted score distribution. Mean ΔF_1 is CAVET-AD minus fixed-rule F1. Confidence intervals are two-sided seed-level Student-t 95% intervals, and Holm p values correct the four paired shift-level hypotheses.

Shift	Mean KS	Mean r_{tail}	Fixed F_1	CAVET F_1	Mean ΔF_1 [95% CI]	W/T/L	Holm p	Interpretation
0.00	0.0000	0.0800	0.7191	0.7834	$0.0642\ [0.0513, 0.0771]$	10/0/0	0.0078	CAVET significantly favorable
0.25	0.1461	0.0921	0.7338	0.7797	$0.0459\ [0.0323, 0.0596]$	10/0/0	0.0078	CAVET significantly favorable
0.50	0.3837	0.1212	0.7423	0.7334	$-0.0089\ [-0.0234, 0.0055]$	4/0/6	0.2754	No significant difference
1.00	0.7463	0.3390	0.7852	0.2750	$-0.5102\ [-0.5567, -0.4638]$	0/0/10	0.0078	Fixed rule significantly favorable

The fitted reference-tail quantities are highly stable across confirmatory seeds. CreditCard2013 is effectively invariant, while CICIoT2023 also shows negligible across-seed variation. CICIDS2018 exhibits only minor variation. NSL-KDD has the largest relative variability among the four datasets, particularly in ξ, but the absolute variation remains small ($\mathrm{SD}(\xi) = 0.001645$ and $\mathrm{SD}(\sigma) = 0.001502$).

The empirical reference exceedance probability is also essentially equal to the value implied by the frozen EVT anchor,

$$1 - q = 0.05,$$

for every dataset. Mean p_u ranges only from approximately 0.049999 to 0.050015 across the four datasets.

Thus, the markedly different confirmatory operating behaviors cannot plausibly be attributed to seed-specific instability of the reference-tail EVT fit. In particular, the exact no-effect behavior of CICIDS2018 should not be interpreted as EVT-fitting instability; its fitted tail model is itself stable across the frozen confirmatory seeds.

G.5 Controlled Covariate-Shift Stress Test

We evaluated whether r_{tail} remains an informative adaptation signal when deployment features change while anomaly prevalence remains fixed. The experiment uses CreditCard2013 with $\pi = 0.10$ and the ten frozen confirmatory seeds. For each seed, deployment observations were translated along the leading principal component estimated from the detector-training partition. Shift magnitudes were

$$\delta \in \{0, 0.25, 0.50, 1.00\}.$$

The detector-training, nominal-reference, and evaluation partitions were held fixed. The detector, EVT reference model, and CAVET-AD configuration were not refitted in response to the deployment shift. Ground-truth labels were used only for controlled mixture construction and retrospective evaluation.

Table 21 shows three regimes. First, the unshifted and mildly shifted conditions preserve a significant CAVET-AD advantage. Second, at shift 0.50, the advantage disappears without a statistically significant reversal. Third, at shift 1.00, the mean score-distribution KS distance reaches 0.7463 and mean r_{tail} increases to 0.3390. This raises the mean adaptive budget to 0.0774 and the CAVET-AD predicted-positive rate to 0.6378, despite the true prevalence remaining fixed at 0.10. CAVET-AD F1 therefore decreases to 0.2750, compared with 0.7852 for the fixed rule.

This result does not contradict the confirmatory prevalence-shift findings. Instead, it identifies their scope: r_{tail} measures excess reference-tail occupancy, but cannot determine whether that occupancy

was caused by increased anomaly prevalence, benign covariate drift, or another change in score geometry. Consequently, CAVET-AD tolerates mild controlled covariate displacement but requires an external drift safeguard or abstention mechanism when large covariate changes are plausible.

Gradual consecutive-batch replay. To evaluate accumulating displacement rather than four isolated jumps, we replayed the same seed-specific CreditCard2013 evaluation membership through 21 consecutive synthetic batches,

$$\delta_b = 0.05(b-1), \qquad b = 1, \ldots, 21,$$

while retaining the fixed prevalence $\pi = 0.10$. The detector, preprocessing transformation, training-derived PCA direction, reference ECDF, EVT fit, and CAVET-AD parameters were fitted once per frozen seed and were not updated across batches. Holding membership fixed within each seed isolates the imposed covariate displacement from batch-composition variation. This is therefore a controlled paired replay, not prospective natural-temporal validation.

Mean r_{tail} increased from 0.0800 (95% seed-level Student-t CI $[0.0779, 0.0821]$) in batch 1 to 0.3390 ($[0.2981, 0.3798]$) in batch 21. The dual-criterion KS safeguard first activated at a median shift of 0.65, with a seed-level range of 0.55–0.70. It activated for $1/10$, $2/10$, $8/10$, and $10/10$ seeds at shifts 0.55, 0.60, 0.65, and 0.70, respectively, and remained active in every later batch after first activation for every seed. Thus, the conservative safeguard does not intervene at the 0.50 performance-crossover regime, but it consistently disables adaptation as severe displacement develops. At $\delta = 1.00$, mean raw CAVET-AD F_1 was 0.2750, whereas the guarded policy retained the fixed rule and achieved mean $F_1 = 0.7852$. The four original anchor conditions reproduced Table 21 to numerical precision.

G.6 Full-Protocol ECOD Replication and Supplementary LOF Transfer

To assess whether the CAVET-AD calibration mechanism transfers beyond Isolation Forest, we evaluated ECOD (Li et al., 2023) and Local Outlier Factor (LOF) (Breunig et al., 2000) after the original confirmatory conclusions and configuration had been frozen. ECOD contains the complete $4 \times 10 \times 9$ condition-level protocol and is treated as the full-protocol second-detector replication in Section 5.5. LOF is retained as a supplementary detector-transfer analysis.

The transfer experiment retained the four confirmatory datasets, the ten authoritative seeds

$$\{7, 11, 13, 21, 23, 37, 51, 55, 73, 97\},$$

and the controlled prevalence grid

$$\pi \in \{0, 0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.60, 0.80\}.$$

The CAVET-AD calibration configuration was also retained unchanged:

$$q = 0.95, \qquad b_0 = 0.01, \qquad \kappa = 0.20, \qquad b_{\max} = 0.10.$$

For every dataset and seed, ECOD and LOF were fitted only on the detector-training partition. Reference-CDF estimation and EVT fitting remained restricted to the nominal reference partition, and controlled prevalence mixtures were constructed only from the held-out evaluation partition. Deterministic bounded subsampling was applied before detector fitting, with the same sampled rows reused across transfer detectors. Neither ground-truth evaluation labels nor true prevalence were supplied to the deployed CAVET-AD rule. No transfer detector or CAVET parameter was selected, removed, or modified on the basis of transfer performance.

CICIDS2018 detector-ranking audit. To separate detector-ranking limitations from threshold-calibration effects, we additionally examined ROC–AUC and average precision computed from the continuous detector scores on the eight positive-prevalence evaluation conditions. These quantities are independent of the fixed and adaptive operating thresholds. Averaging first across the eight prevalence conditions within each seed and then across the ten frozen seeds, ECOD obtained ROC–AUC 0.3422 ± 0.0010 and average precision 0.2627, whereas LOF obtained ROC–AUC 0.9405 ± 0.0102 and average precision 0.7809. Together with the previously reported near-chance Isolation Forest ranking (ROC–AUC 0.516), these results show that the CICIDS2018 limitation is detector-dependent:

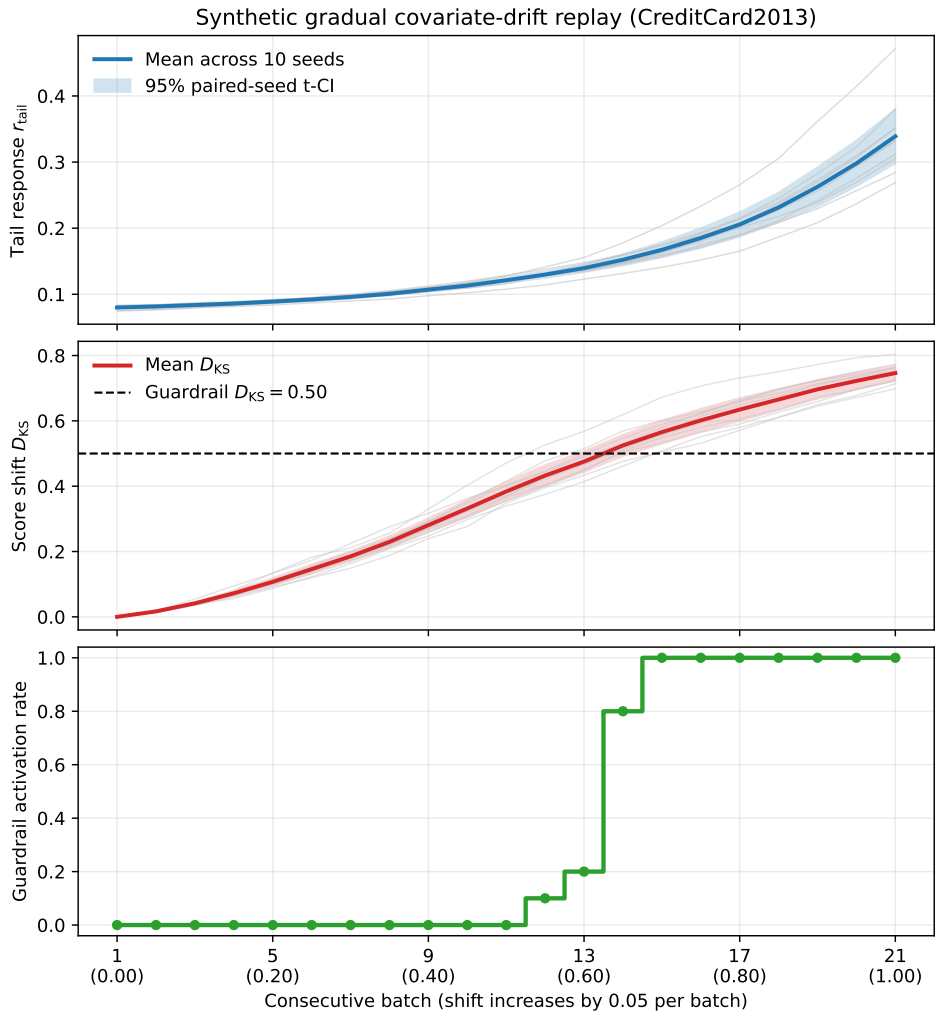


Figure 8: Synthetic gradual covariate-drift replay on CreditCard2013 across ten frozen seeds and 21 consecutive batches. The imposed shift increases by 0.05 per batch while prevalence and evaluation membership remain fixed within each seed. Thin curves show individual seeds, solid curves show seed means, and shaded regions are two-sided 95% seed-level Student-t intervals. The bottom panel reports the fraction of seeds satisfying both $D_{\mathrm{KS}} \geq 0.50$ and $p_{\mathrm{KS}} < 0.01$. This controlled replay evaluates accumulation but is not natural prospective temporal validation.

Table 22: CICIDS2018 detector-score ranking audit over the eight positive-prevalence conditions and ten frozen seeds. ROC–AUC and average precision (AP) are computed from continuous detector scores before applying either operating threshold. The reported ROC–AUC variation is the standard deviation across seed-level means.

Detector	ROC–AUC	Seed SD	AP
Isolation Forest	0.5162	–	–
ECOD	0.3422	0.0010	0.2627
LOF	0.9405	0.0102	0.7809

Isolation Forest and ECOD provide weak or below-chance ranking, while LOF remains strongly informative. Consequently, threshold adaptation cannot rescue the ECOD ranking failure, whereas the favorable LOF result reflects meaningful operating-point adaptation over an informative score ordering.

Table 23: Full-protocol ECOD replication. Confidence intervals are two-sided Student-t 95% intervals over ten paired seed-level units. W/T/L denotes seed-level wins/ties/losses. Holm p values correct the four prespecified ECOD dataset hypotheses.

Dataset	Fixed F_1	CAVET F_1	ΔF_1 [95% CI]	W/T/L; Holm p
CICIDS2018	.0021	.0021	.0001 [.0000,.0001]	7/0/3; .1113
CICIoT2023	.0027	.0027	.0000 [.0000,.0000]	0/10/0; 1.0000
CreditCard2013	.7439	.8055	**.0615 [.0588,.0643]**	**10/0/0; .0078**
NSL-KDD	.0023	.0023	.0000 [.0000,.0000]	0/10/0; 1.0000

Table 24: Post-confirmatory detector-transfer results. Mean ΔF_1 denotes the seed-level paired difference between CAVET-AD and the fixed reference-tail rule after aggregating the eight nonzero prevalence conditions within each of the ten frozen confirmatory seeds. Confidence intervals are two-sided Student-t 95% intervals over those ten paired seed-level effects. W/T/L denotes seed-level wins/ties/losses. Holm p values are adjusted across the eight prespecified dataset–detector comparisons. The corresponding classification precision, recall, and F1 values are reported in Table 25.

Dataset	Detector	Mean ΔF_1	95% CI	W/T/L	Holm p	Sig.	Interpretation
CreditCard2013	ECOD	0.0615	[0.0588, 0.0643]	10/0/0	0.0156	Yes	CAVET favorable
CreditCard2013	LOF	-0.0125	[-0.0163, -0.0088]	0/0/10	0.0156	Yes	Fixed rule favorable
CICIDS2018	ECOD	0.0001	[0.0000, 0.0001]	7/0/3	0.0762	No	Positive, not Holm-significant
CICIDS2018	LOF	0.1809	[0.1407, 0.2210]	10/0/0	0.0156	Yes	CAVET favorable
CICIoT2023	ECOD	0.0000	[0.0000, 0.0000]	0/10/0	1.0000	No	No observed difference
CICIoT2023	LOF	-0.0460	[-0.0480, -0.0440]	0/0/10	0.0156	Yes	Fixed rule favorable
NSL-KDD	ECOD	0.0000	[0.0000, 0.0000]	0/10/0	1.0000	No	No observed difference
NSL-KDD	LOF	0.3252	[0.2211, 0.4293]	10/0/0	0.0156	Yes	CAVET favorable

Scope note. The Isolation Forest value is the held-out evaluation-pool ranking diagnostic from the tail-identifiability audit, whereas the ECOD and LOF entries summarize their detector-transfer evaluations across the frozen seeds.

For inferential analysis, the $\pi = 0$ condition was retained only as an all-normal sanity condition. Across the eight nonzero prevalence conditions, paired effects were first aggregated within each confirmatory seed, yielding ten independent seed-level effects for each dataset–detector combination. Exact two-sided sign-flip tests were then applied across the ten seeds, with Holm correction across the eight prespecified dataset–detector hypotheses.

The dedicated ECOD analysis in Table 23 uses paired Wilcoxon tests with Holm correction over the four prespecified ECOD dataset hypotheses. For continuity with the broader detector-transfer analysis below, the supplementary table retains its original exact sign-flip inference and Holm correction over all eight ECOD/LOF dataset–detector hypotheses. Consequently, its adjusted p-values need not equal those in the dedicated four-hypothesis ECOD analysis; the substantive conclusions are unchanged.

Table 24 shows that the frozen calibration mechanism can be applied to both transfer detectors without detector-specific retuning, although the resulting operating-point effects are heterogeneous.

Within the joint eight-hypothesis ECOD/LOF detector-transfer family, ECOD on CreditCard2013 exhibits a significant positive effect (mean $\Delta F_1 = 0.0615$, Holm-adjusted $p = 0.0156$). The corresponding adjusted value in the dedicated four-dataset ECOD analysis is 0.0078; the difference results solely from the different prespecified multiplicity families. CICIDS2018 exhibits only a very small positive descriptive effect that does not remain significant after Holm correction, while CICIoT2023 and NSL-KDD exhibit no observed paired difference.

For LOF, CAVET-AD is significantly favorable on CICIDS2018 (mean $\Delta F_1 = 0.1809$) and NSL-KDD (mean $\Delta F_1 = 0.3252$). In contrast, the frozen fixed rule is significantly favorable on CreditCard2013 (mean $\Delta F_1 = -0.0125$) and CICIoT2023 (mean $\Delta F_1 = -0.0460$).

Post hoc, the observed LOF sign changes are consistent with dataset-specific precision–recall tradeoffs rather than inconsistent score orientation. On CICIDS2018, adaptation increases recall from 0.4605 to 0.7500, outweighing the precision decrease from 0.8335 to 0.7776. On NSL-KDD, both precision

Table 25: Classification precision, recall, and F1 for the ECOD and LOF detector-transfer experiment. Values are means over the same 80 positive-prevalence seed–condition realizations used for each dataset–detector comparison in Table 24.

		Fixed reference			CAVET-AD		
Dataset	Detector	Precision	Recall	F1	Precision	Recall	F1
CICIDS2018	ECOD	0.0560	0.0012	0.0021	0.0566	0.0013	0.0021
CICIDS2018	LOF	0.8335	0.4605	0.5771	0.7776	0.7500	0.7580
CICIoT2023	ECOD	0.2706	0.0014	0.0027	0.2706	0.0014	0.0027
CICIoT2023	LOF	0.8978	0.9768	0.9270	0.8141	0.9810	0.8810
CreditCard2013	ECOD	0.8726	0.6605	0.7439	0.8001	0.8210	0.8055
CreditCard2013	LOF	0.8811	0.7994	0.8277	0.8002	0.8485	0.8152
NSL-KDD	ECOD	0.2304	0.0012	0.0023	0.2304	0.0012	0.0023
NSL-KDD	LOF	0.6007	0.0737	0.1216	0.6554	0.3774	0.4468

and recall increase, from 0.6007 to 0.6554 and from 0.0737 to 0.3774, respectively. Conversely, on CreditCard2013, the recall increase from 0.7994 to 0.8485 does not compensate for the precision decrease from 0.8811 to 0.8002. On CICIoT2023, recall is already near saturation under the fixed rule (0.9768); adaptation raises it only to 0.9810 while reducing precision from 0.8978 to 0.8141. Thus, the direction of the LOF effect depends on whether the additional alerts generated by adaptive budget expansion provide sufficient recall gain to offset their precision cost.

These precision–recall outcomes require labeled evaluation and are therefore observable only retrospectively. This study does not provide a validated ex-ante, label-free diagnostic for predicting the direction of the LOF effect on a new dataset, and no such predictor is currently available within the proposed framework. Consequently, this analysis should be interpreted as a post-hoc explanation rather than a deployment-selection rule.

Thus, the detector-transfer experiment supports operational *transferability of the frozen calibration mechanism*, but not universal detector-agnostic performance improvement. The direction and magnitude of the benefit remain detector- and dataset-dependent. The CICIDS2018 ranking audit further distinguishes calibration failure from detector failure: CAVET-AD cannot repair the weak or below-chance ranking produced there by Isolation Forest and ECOD, while the strongly informative LOF ranking permits a substantial operating-point improvement. Accordingly, CAVET-AD should be interpreted as a threshold-adaptation mechanism conditional on usable detector-score ordering, rather than as a remedy for detector-ranking failure.

G.7 Deployment-Window-Size Sensitivity

Because CAVET-AD estimates the observable tail response from a finite deployment window, we additionally evaluate sensitivity to the deployment window size m. We vary $m \in \{250, 500, 1000, 2500, 5000\}$ across the nonzero prevalence conditions and the ten frozen confirmatory seeds, subject to the number of available samples in each condition. All other components of the frozen protocol are unchanged.

Figure 10 shows that performance is broadly stable across window sizes for CICIoT2023, CreditCard2013, and NSL-KDD once the deployment window contains sufficient observations. Greater variability appears for the smallest windows, particularly at low prevalence, where the number of anomalous and upper-tail observations is necessarily limited. CICIDS2018 remains a boundary regime with near-zero tail response and correspondingly low F1 across window sizes, consistent with the main experiments.

Table 26 summarizes the within-condition variation across window sizes. These results indicate that finite-window estimation introduces measurable sampling variability, especially for small windows, but does not alter the principal dataset-level regimes observed in the confirmatory experiments.

G.8 Interpretation

Taken together, the post-confirmatory analyses characterize both local configuration robustness and transfer of the frozen calibration mechanism. The EVT-anchor and adaptive-budget experiments quantify sensitivity around the prespecified primary configuration, the nonlinear-schedule comparison

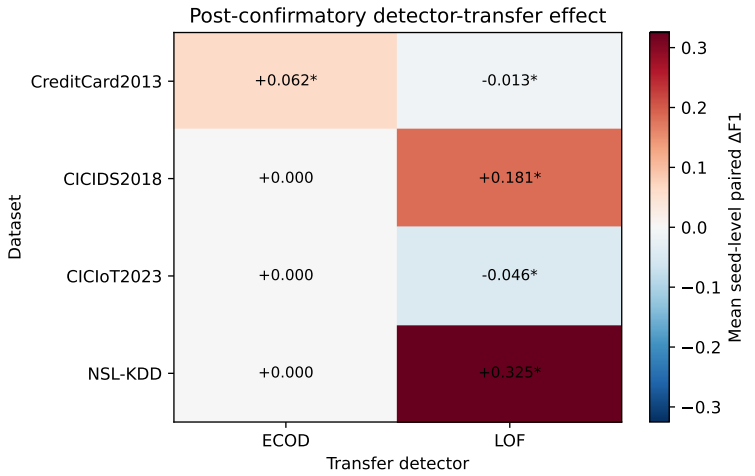


Figure 9: Post-confirmatory detector-transfer effect across ECOD and Local Outlier Factor (LOF). Each cell reports the mean seed-level paired difference $\Delta F_1 = F_{1,\mathrm{CAVET}} - F_{1,\mathrm{fixed}}$ after aggregating the eight nonzero controlled-prevalence conditions within each of the ten frozen confirmatory seeds. Asterisks indicate dataset–detector comparisons that remain significant after Holm correction across the eight prespecified comparisons. Positive values favor CAVET-AD and negative values favor the fixed reference-tail rule. The mixed pattern demonstrates operational detector transfer while also showing that effectiveness remains detector- and dataset-dependent.

Table 26: Deployment-window-size sensitivity summarized across prevalence conditions. F1 range and r_{tail} range denote the within-condition max–min variation across available deployment-window sizes, followed by aggregation across prevalence conditions.

Dataset	$\#\pi$	Mean F_1 range	Median F_1 range	Max F_1 range	Mean r_{tail} range
CICIDS2018	8	0.0079	0.0076	0.0137	0.0027
CICIoT2023	8	0.0365	0.0182	0.1398	0.0078
CreditCard2013	8	0.0375	0.0117	0.1764	0.0033
NSL-KDD	8	0.0365	0.0087	0.1879	0.0079

evaluates the functional form of the budget rule, the EVT-fitting analysis evaluates reference-tail stability, and the detector-transfer experiment evaluates whether the same frozen calibration rule can be applied to additional continuous-score detector families.

None of these analyses was used to select q, b_0, κ, $b_{\max}$, the primary detector, the transfer detectors, the dataset set, the prevalence grid, or any other primary experimental component. The detector-transfer results specifically support operational transferability rather than a claim that CAVET-AD universally improves every detector on every dataset.

Deployment and parameter interpretation. Because the threshold depends on the current batch tail occupancy, the same anomaly score can receive different binary decisions in batches with different score distributions. This behavior is intentional: CAVET-AD adapts the operating point to batch-level tail evidence rather than assigning a globally fixed decision to each score. Applications requiring decisions that are independent of batch composition should instead use a fixed operating rule.

The confirmatory parameters $(q, b_0, \kappa, b_{\max}) = (0.95, 0.01, 0.20, 0.10)$ were fixed before evaluation: q sets the EVT anchor, b_0 the nominal budget, κ the adaptation strength, and $b_{\max}$ its upper bound. Appendix G characterizes robustness around this configuration, but does not establish label-free optimal parameter selection. When $r_{\mathrm{tail}} \approx 0$, moderate changes in κ do not activate adaptation. Adaptive conformal methods (Gibbs & Candès, 2024; Martinez Gil et al., 2026) instead target controlled coverage, while prevalence or contamination estimators (Perini et al., 2022; 2023) could replace the linear budget ramp.

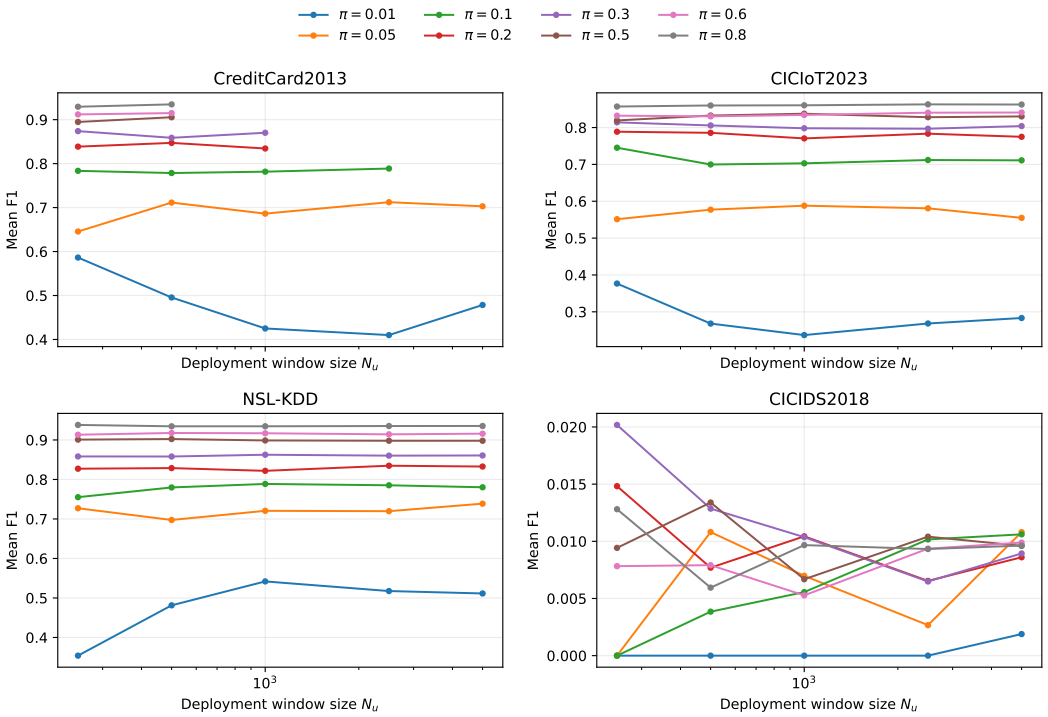


Figure 10: Deployment-window-size sensitivity of CAVET-AD across nonzero prevalence conditions. Curves report mean F1 over the frozen confirmatory seeds for $m \in \{250, 500, 1000, 2500, 5000\}$; unavailable dataset–prevalence–window combinations are omitted. Here N_u in the horizontal-axis labels denotes the realized deployment window size; this quantity corresponds to m in the manuscript notation and should not be confused with the number of EVT exceedances. Performance is generally stable once the deployment window contains sufficient observations, while smaller windows exhibit greater sampling variability.

Although the pipeline accepts continuous detector scores, the strongest cross-detector evidence is the full-protocol ECOD replication in Section 5.5; LOF remains a supplementary transfer analysis (Appendix G.6). The heterogeneous ECOD and LOF effects preclude a claim of universal performance improvement. The controlled covariate-shift stress test establishes one important failure boundary, while joint prevalence–covariate shift, natural temporal deployment, deep detectors, and non-tabular data remain future work.

G.9 Practitioner Pre-Check for Tail-Response Trustworthiness

Before enabling CAVET-AD in deployment, practitioners should verify that the observed excess-tail response is supported by the detector, reference model, and deployment batch rather than treating every large r_{tail} as evidence of prevalence change. We recommend the following pre-check.

Here and throughout the paper, *label-free* refers specifically to the deployed batch-time adaptation rule, which uses only frozen reference calibration and unlabeled deployment scores. It does not refer to end-to-end label-free detector qualification. The optional predeployment detector-ranking audit in Step 1 may use a small historical labeled set; when such labels are unavailable, no claim of label-free detector validation is made.

1. **Detector-ranking audit.** When a small historical labeled audit set is available, verify that the detector provides meaningfully better-than-chance ROC–AUC or average precision. If ranking is near or below chance, threshold adaptation cannot repair the underlying score ordering.

2. **Independent nominal-calibration audit.** Apply the frozen empirical reference CDF to independent nominal observations and inspect the resulting normalized Z-distribution. Large departures from the expected reference behavior indicate that the frozen reference model may no longer be representative.
3. **EVT-support audit.** Require both a nominal reference size of $N_{\mathrm{ref}} \geq 100$ and at least 10 valid exceedances above the frozen anchor $q = 0.95$. The implementation uses GPD maximum likelihood estimation when $n_{\mathrm{exc}} \geq 50$ and a conservative near-exponential fallback when $10 \leq n_{\mathrm{exc}} < 50$. If $N_{\mathrm{ref}} < 100$, $n_{\mathrm{exc}} < 10$, or any fitted parameter is non-finite, adaptation is disabled and the fixed reference-tail rule is retained. Across the 40 frozen dataset–seed reference partitions, N_{ref} ranged from 15,214 to 2,389,875, the exceedance count ranged from 761 to 119,494, and every fit used finite GPD maximum likelihood estimates.
4. **Deployment-window audit.** Compute r_{tail} across reasonable deployment-window sizes or adjacent batches. Large changes caused only by window size indicate unstable tail-occupancy estimation and should prevent adaptation.
5. **Severe score-distribution displacement safeguard.** Let S_0 denote a frozen baseline score sample recorded during an accepted stable deployment or qualification window, and let S_b denote the scores in the current deployment batch. Compute the two-sample Kolmogorov–Smirnov statistic
$$D_{\mathrm{KS},b} = \sup_t \left| \widehat{F}_b(t) - \widehat{F}_0(t) \right|$$
and its two-sided p-value $p_{\mathrm{KS},b}$. For this study-specific post-hoc audit, we flag severe score-distribution displacement only when both
$$D_{\mathrm{KS},b} \geq 0.50 \qquad \text{and} \qquad p_{\mathrm{KS},b} < 0.01.$$
The effect-size requirement prevents very large batches from triggering the safeguard for small but statistically detectable differences. When the safeguard fires, CAVET-AD adaptation is disabled and the frozen fixed reference-tail rule is retained.
As a worked example, for CreditCard2013 seed 7, the observed $(D_{\mathrm{KS}}, p_{\mathrm{KS}})$ pairs at shift levels 0, 0.25, 0.50, and 1.00 were respectively $(0, 1)$, $(0.1506, 4.50 \times 10^{-49})$, $(0.3711, 7.14 \times 10^{-302})$, and $(0.7632, 0$ at machine precision$)$. Thus, the first three conditions passed the guardrail, whereas the severe-shift condition disabled adaptation. Across the ten frozen seeds, the corresponding flag counts were $0/10$, $0/10$, $0/10$, and $10/10$. These results are consistent with the performance boundary in Table 21.
This score-space diagnostic is a conservative abstention safeguard, not a causal classifier of drift: a flagged displacement may arise from covariate shift, prevalence shift, concept drift, or another change. The 0.50 cutoff is therefore not claimed as universal and should be prospectively recalibrated for a new application.

CAVET-AD should activate adaptation only when detector ranking is usable, independent nominal calibration is acceptable, EVT-tail support is adequate, the tail response is stable, and the severe-displacement safeguard does not fire. If any check fails, the operational fallback is to retain the fixed reference-tail rule or abstain pending review. The deployment-time checks use only frozen calibration information and unlabeled score batches; historical labels may be used only for the optional offline detector-ranking audit in Step 1.

H Boundary Regimes and Diagnostic Interpretation

The confirmatory experiment reveals several non-uniform operating regimes that are important for interpreting CAVET-AD.

H.1 CreditCard2013: Lowest-Prevalence Uncertainty

At the lowest tested nonzero prevalence, $\pi = 0.01$, the paired F1 difference is slightly negative ($\Delta F_1 = -0.0073$), but it does not remain significant after Holm correction ($p_{\mathrm{Holm}} = 0.0586$).

This result shows that the positive aggregate effect on CreditCard2013 should not be interpreted as uniform improvement in the most extreme rare-event regime.

Table 27: CreditCard2013 evaluation at the natural prevalence $\pi = 0.00173$ over ten paired frozen seeds. Differences are CAVET-AD minus the fixed reference-tail rule.

Metric	Fixed	CAVET-AD	Difference
Precision	0.09696	0.09613	-0.00083
Recall	0.62023	0.62601	$+0.00578$
F1	0.16765	0.16658	-0.00107
Predicted-positive rate	0.01111	0.01133	$+0.00022$
Operating budget	0.01000	0.01023	$+0.00023$

H.2 CREDITCARD2013: NATURAL-PREVALENCE EVALUATION

We additionally evaluated the frozen Isolation Forest implementation at the realized CreditCard2013 evaluation-pool prevalence of $\pi = 0.00173$ (0.173%). The experiment used the ten frozen seeds $\{7, 11, 13, 21, 23, 37, 51, 55, 73, 97\}$ and 100,000 evaluation observations per seed, comprising approximately 99,827 normal and 173 anomalous observations. The detector, empirical reference CDF, EVT model, and CAVET-AD parameters remained frozen; no detector training, refitting, hyperparameter selection, or threshold tuning was performed using the natural-prevalence evaluation labels.

Before evaluating the natural-prevalence condition, the recovered implementation was validated against an archived controlled CreditCard2013 condition. The reproduced threshold, precision, recall, F1, and predicted-positive-rate values matched their saved values for both the fixed-reference and CAVET-AD rules.

For F1, the mean paired difference was -0.00107, with an unadjusted two-sided paired-seed Student-t 95% interval of $[-0.00212, -0.00003]$. The raw two-sided paired Wilcoxon result for F1 was $p \approx 0.028$. After Holm correction across the six prespecified natural-prevalence comparisons (precision, recall, F1, predicted-positive rate, operating budget, and threshold), the F1 result was not significant ($p_{\mathrm{Holm}} \approx 0.108$). We therefore do not conclude a multiplicity-adjusted significant F1 improvement or deterioration.

CAVET-AD exchanged a small increase in recall for a small decrease in precision while changing the operating budget by only approximately 2.3×10^{-4}. The favorable aggregate CreditCard2013 result from the controlled prevalence sweep therefore should not be extrapolated to the dataset's substantially rarer natural prevalence. Instead, the result supports the conditional interpretation that adaptation remains minimal when the deployment condition provides limited actionable excess-tail evidence.

H.3 NSL-KDD: FIXED-TO-ADAPTIVE CROSSOVER

NSL-KDD exhibits a clear prevalence-dependent crossover. The frozen fixed reference-tail operating rule is significantly favorable at $\pi = 0.01$, 0.05, and 0.10. The $\pi = 0.20$ condition is statistically unresolved, whereas CAVET-AD becomes significantly favorable from $\pi = 0.30$ onward.

This result demonstrates that the preferred operating rule can depend on the prevalence regime rather than being uniformly dominated by one thresholding strategy.

H.4 CICIDS2018: NON-IDENTIFIABLE BOUNDARY CASE

CICIDS2018 produces an exact aggregate paired F1 difference of zero and no multiplicity-adjusted CAVET-favorable prevalence condition. The retrospective tail-identifiability diagnostic is negative, indicating that the anomalous observations do not produce the positive excess reference-tail response required for useful adaptation.

This case is interpreted as absence of usable positive reference-tail information for adaptation, not as evidence of algorithmic failure or unstable EVT fitting.

Table 28: Audited software and experimental environment for the frozen CAVET-AD confirmatory analysis. Versions were extracted programmatically from the active execution environment. The listed experimental settings correspond to the frozen confirmatory configuration and were not reconstructed from manuscript values.

Reproducibility item	Frozen specification
Python	3.9.18
NumPy / pandas	2.0.2 / 2.2.2
SciPy / scikit-learn	1.13.1 / 1.6.1
Matplotlib / PyArrow / joblib	3.9.4 / 21.0.0 / 1.4.2
Platform	Linux, x86_64, 32 CPUs
Frozen seeds	7, 11, 13, 21, 23, 37, 51, 55, 73, 97
Prevalence grid	0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.6, 0.8
EVT / support configuration	generalized_pareto; $q = 0.95$; $N_{\mathrm{ref}} \geq 100$; ≥ 10 valid upper-tail exceedances for support audit
Adaptive budget	$b_0 = 0.01$, $\kappa = 0.20$, $b_{\max} = 0.10$

H.5 Diagnostic Interpretation

Taken together, these cases reinforce the conditional interpretation of CAVET-AD: adaptive thresholding is beneficial when the observable reference-tail response contains sufficient information for adaptation. Weak, transitional, or directionally incompatible tail-response regimes can reduce, eliminate, or temporarily reverse the operating-point benefit.

I Reproducibility & Open-Science Checklist

I.1 Objective

To support transparency and replicability, the CAVET-AD experiments use fixed random seeds, publicly available datasets, and open-source software dependencies. This section summarizes the data sources, frozen configurations, software environment, and execution controls underlying the reported results.

I.2 Datasets

All datasets used in this study are publicly available and accessed via their canonical research repositories.

The supplementary code separates the frozen primary experiment from the post-confirmatory robustness and diagnostic analyses reported in this appendix. Randomness is controlled by explicit prespecified seeds, with the frozen confirmatory analysis using the ten-seed set $\{7, 11, 13, 21, 23, 37, 51, 55, 73, 97\}$.

I.3 Experimental Environment

The environment manifest was generated programmatically from the active execution environment and the frozen confirmatory protocol. The audit also indexed 172 result artifacts under the experiment output tree and verified the presence of the frozen partition, paired-result, EVT-sensitivity, budget-sensitivity, prevalence-inference, and evidence-registry objects. No detector, partition, reference model, EVT fit, controlled prevalence mixture, confidence-interval procedure, sign-flip test, or multiplicity correction was recomputed while constructing the reproducibility manifest.

Version-control metadata could not be recovered from the active notebook working directory and are therefore reported as unavailable rather than inferred. This limitation concerns repository provenance only and does not alter the frozen experimental objects or reported numerical results.

Environment validation. The recorded environment corresponds to the active Linux execution environment used for the audited analysis. Software versions, frozen experimental settings, and

Table 29: Principal audited artifacts from the frozen primary CAVET-AD consolidation. All listed artifacts were verified as non-empty and assigned SHA-256 checksums. Additional post-confirmatory artifacts are documented separately from this primary closure inventory.

Artifact	Audited output
Independent nominal calibration	`independent_nominal_qq_all_datasets_seed7.png`
Exp(1) calibration diagnostic	`independent_nominal_exp1_gof_seed7.csv`
Frozen confirmatory design	`table_s2_frozen_confirmatory_design_compact.csv`
κ provenance	`kappa_provenance_summary.csv`
EVT-anchor sensitivity	`figure_s2_evt_anchor_sensitivity.png`
Adaptive-budget sensitivity	`figure_s3_adaptive_budget_sensitivity.png`
Detector-transfer analysis	`figure_s4_detector_transfer.png`
Detector-transfer summary	`table_s7_detector_transfer_compact.csv`
EVT-fitting stability	`evt_fitting_stability_compact.csv`
Tail identifiability	`tail_identifiability_compact.csv`
Prevalence-resolved inference	`prevalence_inference_table_s6_compact.csv`
Partition/provenance	`partition_provenance_table_compact.csv`
Environment/reproducibility	`reproducibility_manifest_compact.csv`
Machine-readable environment	`environment_manifest.json`

artifact availability were captured programmatically. No claim of bitwise identity across different operating systems or hardware platforms is made by this audit.

I.4 Reproducibility Protocol

1. Setup

- Extract the anonymous supplementary release package `CAVET-AD-ICLR`.
- Install dependencies from `requirements.txt`.

2. Data Acquisition Dataset acquisition, canonical sources, preprocessing conventions, label mappings, and partition construction are documented in Appendix E. The datasets are not redistributed with the anonymous supplementary materials.

3. Experiment Execution An anonymized implementation and experimental materials accompany this submission. Post-confirmatory robustness and diagnostic analyses are executed separately without modifying the frozen primary configuration or confirmatory results. Generated manuscript artifacts are stored under `results/`.

4. Generated Artifacts Experiment outputs include calibration diagnostics, confirmatory result tables, post-confirmatory robustness artifacts, statistical summaries, provenance audits, and reproducibility manifests. The frozen primary consolidation contained 85 indexed files, and its manuscript-facing integrity manifest identified 14 required artifacts; all 14 were verified to exist, to be non-empty, and to have SHA-256 checksums. Subsequent post-confirmatory analyses generated additional manuscript-linked artifacts, including full-protocol ECOD replication, detector transfer, group-disjoint robustness, stronger-baseline comparison, component ablation, mechanism analysis, leave-one-dataset-out applicability, controlled covariate-shift, window-size sensitivity, and prespecified nonlinear-budget comparison outputs, without modifying the frozen primary consolidation or confirmatory results. The held-out applicability artifacts are stored under `results/iclr_postconfirmatory/prospective_rtail_gate_lodo_v1/` and include the condition-level predictions, fold summaries, leakage audit, statistical tests, locked experiment summary, and SHA-256 manifest. Table 29 lists the principal manuscript-linked outputs rather than every intermediate file.

5. Reproduction Verification The frozen reproducibility audit verifies the confirmatory dataset set, ten-seed set, controlled prevalence grid, EVT anchor, adaptive-budget configuration, partition provenance, and availability of the principal frozen result objects. The audit additionally records the active software environment and generated artifact inventory without rerunning any experimental or statistical procedure.

Table 30: Open-science reproducibility checklist for CAVET-AD.

Criterion	Status	Description
Open Datasets	✓	All datasets publicly available
Supplementary Code	✓	Anonymized implementation and experimental materials provided with the submission
Random Seeds Fixed	✓	Frozen 10-seed confirmatory set; development seed 42 excluded
Operating Configuration	✓	Frozen primary configuration and post-confirmatory sensitivity grids documented
Evaluation Metrics Defined	✓	F1, precision, recall, predicted-positive rate; AUROC/AP as ranking diagnostics
Statistical Tests	✓	Paired sign-flip and Wilcoxon signed-rank inference with Holm multiplicity correction
Post-confirmatory robustness	✓	EVT-anchor, adaptive-budget, nonlinear-schedule, EVT-fitting, detector-transfer, group-disjoint, and covariate-shift analyses documented
Comparator and ablation analyses	✓	Stronger reference-only baselines and component ablations documented
Mechanism diagnostics	✓	Tail-identifiability and observable tail-response/adaptation-gain analyses documented
Provenance audit	✓	Partition, environment, artifact, and final closure audits completed
Implementation Materials	✓	Anonymized notebooks, source modules, dependencies, and documentation provided
Version Control / DOI	pending	Zenodo submission planned

The leave-one-dataset-out applicability analysis additionally verifies that each cutoff was fitted using exactly three development datasets and that the held-out dataset was absent from cutoff selection. Its archived output contains 320 condition-level predictions, 40 dataset–seed inference units, four fold summaries, and a passing four-fold leakage audit.

I.5 Open-Science Checklist

This subsection summarizes the reproducibility controls applied to the CAVET-AD study. Table 30 records dataset availability, the frozen random-seed design, documented operating configurations, evaluation metrics, multiplicity-adjusted paired inference, artifact generation, and the accompanying anonymized implementation. The accompanying environment manifest records the software versions and frozen experimental specification used in the reported analysis. Repository-level version-control provenance was not available from the active notebook environment and is therefore not inferred.

Final closure audit. A final read-only consolidation audit verified that the reported analysis retains the four frozen datasets, the authoritative confirmatory seed set $\{7, 11, 13, 21, 23, 37, 51, 55, 73, 97\}$, the nine prespecified prevalence conditions, the primary EVT anchor $q = 0.95$, and the adaptive-budget configuration $(b_0, \kappa, b_{\max}) = (0.01, 0.20, 0.10)$. The audit further verified nominal-only detector-training and reference partitions, exclusion of the held-out evaluation pool from detector, reference-CDF, and EVT fitting, and the absence of labels or true prevalence from the deployed adaptation rule.

All 14 manuscript-required artifacts included in the frozen primary consolidation were verified to exist as non-empty files and were assigned SHA-256 checksums. The later post-confirmatory artifacts, including detector-transfer, group-disjoint robustness, stronger-baseline comparison, component-ablation, and mechanism-analysis outputs, were generated after the frozen primary closure and were therefore not part of that earlier closure manifest.

No confirmatory experiment, sensitivity experiment, confidence-interval procedure, paired sign-flip analysis, or multiplicity correction was recomputed during this closure audit, and no post-confirmatory robustness result was used to alter the frozen primary configuration.

I.6 License and Citation

An anonymized implementation and supplementary materials accompany this submission.

Summary The frozen confirmatory analysis is accompanied by audited dataset, row-index partition, seed, prevalence-grid, operating-configuration, statistical, software-environment, and artifact provenance. The final primary closure audit verified internal consistency of the frozen experimental design and all 14 primary manuscript-required artifacts without recomputing any experimental or inferential procedure. Subsequent post-confirmatory artifacts document the group-disjoint robustness study, stronger comparator evaluation, component ablation, detector transfer, and mechanism analyses without altering the frozen primary configuration or confirmatory results. Repository-level version-control provenance and public code-release metadata remain pending and are not inferred.