SAME MAGNITUDE, DIFFERENT DIRECTION: SIGNED-ORIENTATION RESPONSE IN TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern time-series anomaly detectors respond very differently to signed cross-channel relations, and the corresponding score component can help or harm detection on natural data. We establish these differences using a partial channel sign flip that preserves every coordinatewise magnitude. Across nine released models on 65 multivariate views, responses range from near chance to strong separation; longer training changes them unevenly. An exact even–odd decomposition isolates the flip-sensitive component of the same trained score, linking the controlled comparison to a natural-detection ablation. At 50 epochs, removing this component changes magnitude-matched AUC by more than one percentage point in 19 of 54 model–dataset combinations and by as much as 14.37 points. Yet Anomaly Transformer and GDN, only 1.00 native-AUC point apart on a common panel, differ by 11.36 points in response strength. Similar native performance is therefore insufficient evidence of similar signed response: this behavior and its effect on detection require direct measurement.

1 INTRODUCTION

Two channels can have the same magnitudes and opposite relations. For any temporal pattern u, the windows (u, u) and $(u, -u)$ have identical coordinatewise magnitudes, but their cross-products are u_t^2 and $-u_t^2$. A magnitude-only representation gives these patterns the same description; aggregation can also cancel opposing contributions, as $a + (-a) = 0$. We call this distinction *signed orientation*.

Relationships between channels already motivate graph-based, probabilistic, and causal approaches to multivariate time-series anomaly detection (TSAD) (Deng & Hooi, 2021; Chen et al., 2022; Liu et al., 2025). For these models, the input distinction raises a question about the final trained score: *how strongly does it distinguish the two directions at fixed magnitude, and when does that response help natural anomaly detection?*

The answer differs sharply across trained models. Across nine released implementations on 65 multivariate views, some scores strongly separate original and sign-flipped inputs while others remain close to chance. On a separate common panel, extending training from 5 to 50 epochs strengthens several responses but leaves others almost unchanged. The studies cover industrial control, spacecraft telemetry, and server monitoring; the common panel contains 56 views for paired natural-detection and training-budget comparisons.

To establish these differences, we flip a channel group chosen from normal training data. Every coordinatewise magnitude is preserved, while signed products across the flipped/unflipped partition reverse. For the same trained score, A_T measures rank separation between original and flipped normal windows, and $\mathcal{C}_T = |A_T - 0.5|$ records its strength irrespective of direction. Exact score invariance gives $A_T = 0.5$. Across the nine models, mean response strength ranges from 0.011 to 0.233.

The corresponding score component also changes natural detection. The same two score evaluations split each trained score into a flip-invariant part and a part that reverses under the flip (Figure 1). Removing the latter changes magnitude-matched AUC by more than one point in $19/54$ model–dataset combinations at 50 epochs. In the pre-specified four-model subgroup, the mean absolute

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

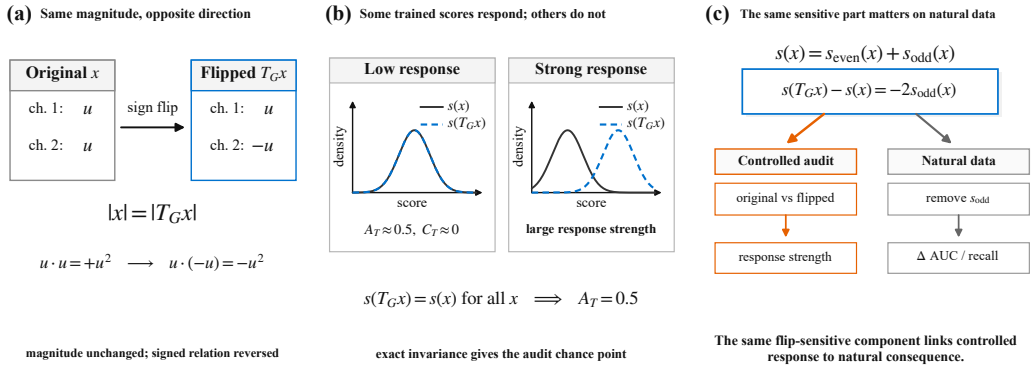

Figure 1: Signed orientation in a trained score. (a) The two-channel example has identical coordinatewise magnitudes and opposite signed products. (b) Original and flipped windows can produce overlapping or separated score distributions; exact score invariance gives $A_T = 0.5$. (c) The flip-sensitive part determines the pointwise score change and is removed to measure its effect on natural detection. The distributions are schematic.

change in native AUC is 2.63 points. The component helps detection on some datasets and harms it on others, even within the same model family. Strong expression of the signed distinction and beneficial use of it are separate properties.

Native AUC does not resolve these response differences. On the common panel, Anomaly Transformer and GDN differ by only 1.00 native-AUC point but by 11.36 points in response strength. We formalize this mismatch through capability certification: whether nearby benchmark scores constrain a specified response to be nearby. The released-model pairs violate that requirement at the stated tolerances. A separate fixed-candidate selection study provides supporting evidence: selection by validation response and selection by native AUC lead to different held-out responses. Similar detection performance is therefore insufficient evidence that two trained scores respond to this distinction with similar strength.

Main findings.

- **Signed response varies across trained models.** The same equal-magnitude contrast produces markedly different responses across nine released implementations; longer training changes their strengths unevenly.
- **The flip-sensitive component changes natural detection.** Removing the flip-sensitive component can improve or worsen AUC and recall on natural data.
- **Native performance leaves the response unresolved.** Near-equal native AUCs coexist with large response differences.

2 Related Work

TSAD models and relationships between channels. TSAD methods include statistical, reconstruction, forecasting, graph, attention, and frequency-based approaches (Chandola et al., 2009; Pang et al., 2021; Zamanzadeh Darban et al., 2024). Our released-model panel covers USAD, GDN, Anomaly Transformer, TimesNet, ModernTCN, DCdetector, CATCH, ScatterAD, and Time-RCD (Audibert et al., 2020; Deng & Hooi, 2021; Xu et al., 2022; Wu et al., 2023; Luo & Wang, 2024; Yang et al., 2023; Wu et al., 2025; Yin et al., 2025; Lan et al., 2026). Several methods explicitly model relationships between variables: GDN learns sensor relationships, variational graph models represent spatial–temporal correlations, and causal approaches examine Granger-causal structure (Deng & Hooi, 2021; Chen et al., 2022; Liu et al., 2025). Here the object of study is the final score's signed-orientation response across model families. The controlled contrast measures that response

directly; the score decomposition tests whether its flip-sensitive component improves detection on natural data.

Evaluation and behavioral testing. TSAD comparisons depend on labels, point adjustment, metrics, data composition, post-processing, and model-selection protocols (Campos et al., 2016; Kim et al., 2022; Schmidl et al., 2022; Paparrizos et al., 2022; Liu & Paparrizos, 2024; Qiu et al., 2025; Zhou et al., 2026). Underspecification and hidden stratification show more broadly how similar aggregate performance can coexist with different model behavior (D'Amour et al., 2022; Oakden-Rayner et al., 2020). ImageNet-C, WILDS, and CheckList likewise test behaviors that a single aggregate score can obscure (Hendrycks & Dietterich, 2019; Koh et al., 2021; Ribeiro et al., 2020). Our study characterizes how strongly trained scores express signed orientation, what the corresponding component changes on natural data, and whether native AUC resolves that response.

3 Signed Orientation in a Trained Score

3.1 Hold magnitude fixed and change direction

Let $x \in \mathbb{R}^{L\times d}$ be a residual window with L time steps and d channels. Choose a nonempty proper subset of channels G from normal training data and define

$$(T_G x)_{t,j} = \begin{cases} -x_{t,j}, & j \in G, \\ x_{t,j}, & j \notin G. \end{cases} \tag{1}$$

Flipping twice returns the original window: $T_G^2 = I$. Every coordinatewise magnitude is unchanged. For a flipped channel $j \in G$ and an unflipped channel $k \notin G$,

$$(T_G x)_{t,j}(T_G x)_{t,k} = -x_{t,j}x_{t,k}. \tag{2}$$

Products between two flipped channels or two unflipped channels stay unchanged. Thus the signed products across the partition reverse exactly, with no coordinatewise magnitude change. An asymmetric signed marginal can also change under the flip; Appendix B tests this alternative explanation with independent-channel, invariant-score, channel-selection, and shuffle controls.

3.2 Measure the score response

For a fixed trained score s, compute AUC between original and flipped normal windows, treating the flipped windows as the positive class:

$$A_T(s) = \mathrm{AUC}\big(s(X), s(T_G X)\big), \qquad \mathcal{C}_T(s) = \left| A_T(s) - \frac{1}{2} \right|. \tag{3}$$

Here AUC compares the two score distributions: for independent normal windows X, X', it equals $\Pr[s(T_G X') > s(X)]$ plus half the tie probability. $A_T > 0.5$ means flipped windows tend to receive higher anomaly scores; $A_T < 0.5$ means the ordering is reversed. We call $\mathcal{C}_T$ the *response strength*. For example, $A_T = 0.8$ and $A_T = 0.2$ both give $\mathcal{C}_T = 0.3$: equally strong rank separation in opposite directions. One response point is 0.01 on the $[0, 0.5]$ scale. AUC and recall differences are reported in percentage points.

An exact reference point. If $s(T_G x) = s(x)$ for every x, the two score distributions are identical, so $A_T = 0.5$ and $\mathcal{C}_T = 0$. This gives an exact chance reference for flip-invariant scores. Appendix A states the result for invariant representations and gives the proof.

3.3 Separate response from its effect on detection

The same two score evaluations let us split s into a part that stays unchanged under the flip and a part that reverses:

$$s_{\mathrm{even}}(x) = \frac{s(x) + s(T_G x)}{2}, \qquad s_{\mathrm{odd}}(x) = \frac{s(x) - s(T_G x)}{2}. \tag{4}$$

We refer to these as the *flip-invariant* and *flip-sensitive* parts, respectively; they are the even and odd parts under T_G. They sum to the original score, and

$$s(T_G x) - s(x) = -2s_{\text{odd}}(x). \tag{5}$$

The score change caused by the flip is therefore determined pointwise by the part we remove. $\mathcal{C}_T$ summarizes rank separation on normal windows; s_{odd} gives the pointwise flip-sensitive contribution. We evaluate the latter on labeled natural data by comparing the full score with s_{even}.

For natural labels y, define the full-minus-invariant AUC difference

$$D_{\text{nat}} = \text{AUC}(y, s) - \text{AUC}(y, s_{\text{even}}). \tag{6}$$

A positive value means the flip-sensitive part helps detection; a negative value means removing it improves detection. $|D_{\text{nat}}|$ records the size of the change. This separates two questions: how strongly the model responds to the flip, and whether that response helps detection on a given labeled dataset.

4 Experimental Setup

Models and data. We use frozen public revisions of the nine models listed above, retaining each implementation's architecture, training objective, and scoring rule (Appendix C). All receive the same residualization, calibration, and window construction. Throughout, *native AUC* means detection AUC on labeled, untransformed windows under this shared protocol. A *view* is a data file or a fixed evaluation segment. The primary response study contains 65 multivariate views; three of the original 68 candidates are univariate and cannot support a channel partition. A separate common panel contains 56 views from six datasets: HAI 13, MSL 7, PSM 1, SMAP 13, SMD 21, and SWaT 1. The sources include industrial-control data (Shin et al., 2020; Mathur & Tippenhauer, 2016), spacecraft telemetry (Hundman et al., 2018), and server-machine monitoring (Su et al., 2019). This panel supports paired comparisons of native AUC, natural score changes, and training budgets.

Preprocessing and training. We use VAR(1) residuals with ridge 10^{-4}, a train-only empirical-CDF rank-Gaussian calibration, windows of length 50 and stride 10, and a top-correlation group of size $|G| = \min(8, d-1)$ chosen from training data. Eight trainable models use seeds $0, 1, 2$; Time-RCD uses its released zero-shot checkpoint. The protocol specifies 5 epochs for the primary response study and a longer 50-epoch budget for comparison, with all other settings fixed. We form $\mathcal{C}_T = |A_T - 0.5|$ separately for each seed and view, then average across seeds, views within each dataset, and finally datasets with equal weight. Released-model confidence intervals use 50,000 hierarchical-bootstrap replicates, resampling datasets, views, and seeds in that order.

Natural detection. We compare s and s_{even} on identical labeled positions. Alongside native AUC, we report AUC after matching attack and normal windows on magnitude. We use nearest matching without replacement based on the window magnitude $m(w) = L^{-1}\sum_t \|w_t\|_2^2$, which is invariant under the sign flip. The caliper is 0.10 times the interquartile range of $m(w)$ among normal windows. Matching is fixed before detector scores are inspected; an evaluated view is retained only if at least ten attack windows can be matched to normal windows. Recall is measured at the same empirical 5% false-positive rate (FPR), with a separate threshold for each score calibrated on the corresponding evaluation-normal controls. Event recall uses contiguous attack-window clusters. The four-model high-response subgroup—Anomaly Transformer, USAD, GDN, and TimesNet—was fixed from the 5-epoch audit before natural results were inspected; the same four models are used in every bootstrap draw. The full-panel counts summarize the size of the dataset-average effect for each model. We first average the signed AUC change across seeds and views, then take the absolute value of that model–dataset mean and count how many pairs exceed 1, 2, or 5 percentage points. The subgroup summary asks a different question: how large is the effect regardless of direction? For that analysis, we take the absolute AUC change for each seed and view first, then average these magnitudes across seeds, views within each dataset, and datasets with equal weight.

Controls. In the focused checks, exactly sign-invariant scorers give audit AUC 0.500 ± 0.000, while a scorer using independent-channel marginals remains near chance. Alternative channel groups and relational shuffles test the role of channel selection and cross-channel alignment (Appendix B).

Table 1: Response strength on 65 multivariate views. We average seeds within each view, views within each dataset, and datasets equally. Intervals are 95% hierarchical-bootstrap CIs.

Model	$\mathcal{C}_T$	95% CI
USAD	0.233	[0.072, 0.399]
Anomaly Transformer	0.174	[0.121, 0.242]
TimesNet	0.121	[0.033, 0.220]
GDN	0.117	[0.038, 0.238]
ScatterAD	0.030	[0.012, 0.059]
ModernTCN	0.026	[0.002, 0.061]
Time-RCD	0.020	[0.010, 0.031]
CATCH	0.014	[0.002, 0.040]
DCdetector	0.011	[0.003, 0.023]

Model selection. The nine-model comparison uses native AUC and $\mathcal{C}_T$ from the same runs on the common panel. A separate 46-fold, six-dataset study uses 13 fixed candidate methods, including analytic, classical, and neural scores (Appendix E). Its pre-specified primary analysis selects by directional A_T; the post-hoc response-strength analysis uses $\mathcal{C}_T$ from the same archived seed-level outcomes. Selection intervals resample datasets and folds within datasets, with equal dataset weighting and 50,000 bootstrap replicates. A supplementary 17-score comparison measures response differences among scores within specified native-AUC bands (Appendix D).

5 Results

5.1 Signed response differs across models and training budgets

The same signed contrast produces a wide range of responses across the nine released models. On the 65-view panel, response strength ranges from 0.011 for DCdetector to 0.233 for USAD. Anomaly Transformer reaches 0.174, TimesNet 0.121, and GDN 0.117; the other model averages are 0.030 or below (Table 1).

Longer training changes this pattern unevenly. On the common 56-view panel, Anomaly Transformer moves from 0.287 to 0.352, GDN from 0.173 to 0.266, TimesNet from 0.130 to 0.255, and ModernTCN from 0.026 to 0.184 between 5 and 50 epochs. The respective increases are 6.55, 9.29, 12.43, and 15.76 response points. DCdetector remains close to chance, while CATCH and ScatterAD change by less than half a point. The pre-specified four-model subgroup gains 6.86 points on average (95% CI $[4.39, 9.43]$). Increasing the training budget changes signed response within the same model family without eliminating the differences across models.

Response differences persist across channel groups. On the 65-view panel at 5 epochs, seven alternatives to the canonical group retain similar nine-model rankings (Spearman $\rho = 0.917$–0.983; Appendix L). These settings vary group size or select low-correlation or random channels. The same four models occupy the top four positions in all eight settings. The three random-group replicates yield a mean response across models of 0.102–0.107, compared with 0.083 for the canonical group.

5.2 The flip-sensitive component changes natural detection in both directions

The component isolated by the controlled flip also affects detection on labeled natural data. We replace each trained score by s_{even} on the common natural panel, holding the model, labels, and evaluated positions fixed. At 50 epochs, the absolute mean change in magnitude-matched AUC exceeds one point in $19/54$ model–dataset combinations (35.2%), two points in $10/54$ (18.5%), and five points in $6/54$ (11.1%). The full-minus-invariant difference ranges from -3.15 to $+14.37$ AUC points. The flip-sensitive part helps detection on some datasets and harms it on others. For Anomaly Transformer, it improves magnitude-matched AUC on HAI and SMAP but worsens it on MSL, PSM, and SMD (Figure 3). These opposite effects within a model family show why response strength and natural usefulness must be measured separately.

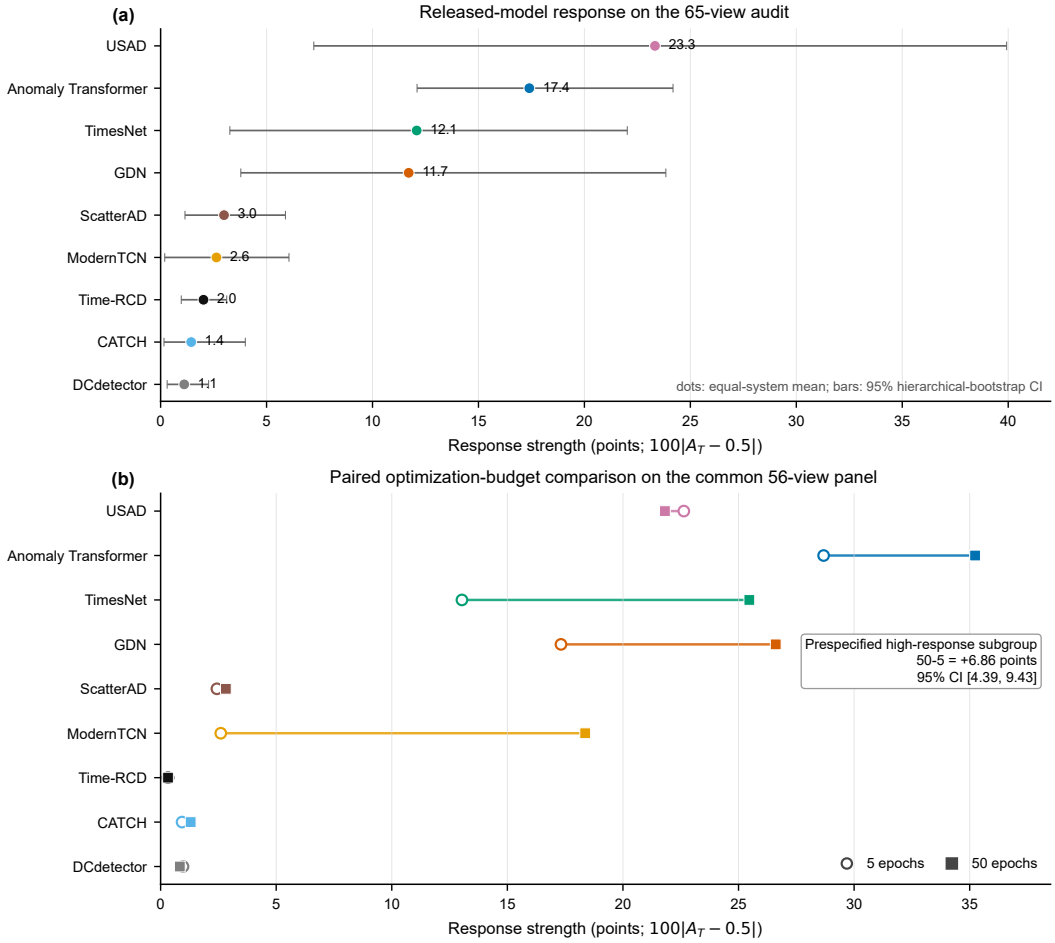

Figure 2: Responses to the same flip differ across models and training budgets. The 65-view response study and the paired 56-view training comparison are shown separately.

Table 2: Natural detection changes after removing the flip-sensitive part, at 50 epochs. Values are mean absolute percentage-point changes in the pre-specified subgroup: Anomaly Transformer, USAD, GDN, and TimesNet. Recall uses a 5% FPR.

Metric	Absolute change	95% CI
Native (unmatched) AUC	2.63	[1.63, 4.32]
Magnitude-matched AUC	3.46	[1.87, 5.48]
Unmatched window recall	4.11	[2.12, 6.68]
Event recall	5.32	[3.03, 8.06]

The effect also appears in native AUC. In the pre-specified four-model subgroup, the mean absolute change in native (unmatched) AUC after removing the flip-sensitive part is 2.63 **points** at 50 epochs (95% CI $[1.63, 4.32]$). The corresponding magnitude-matched change is 3.46 points, and event recall changes by 5.32 points (Table 2). These are changes in detection on the original labeled data, measured using the full score and its invariant part. The mean absolute native-AUC change rises from 1.75 points at 5 epochs to 2.63 at 50 epochs, a paired increase of 0.88 points ($[0.11, 1.84]$).

Larger responses tend to accompany larger natural changes. Across the *nine model averages* on the common panel, the descriptive Pearson correlation between $\mathcal{C}_T$ and the absolute magnitude-matched AUC change is $r \approx 0.82$ at 5 epochs and $r \approx 0.89$ at 50 epochs. Each coefficient summarizes nine models at one budget. For example, the absolute matched-AUC changes at 50 epochs are 5.84

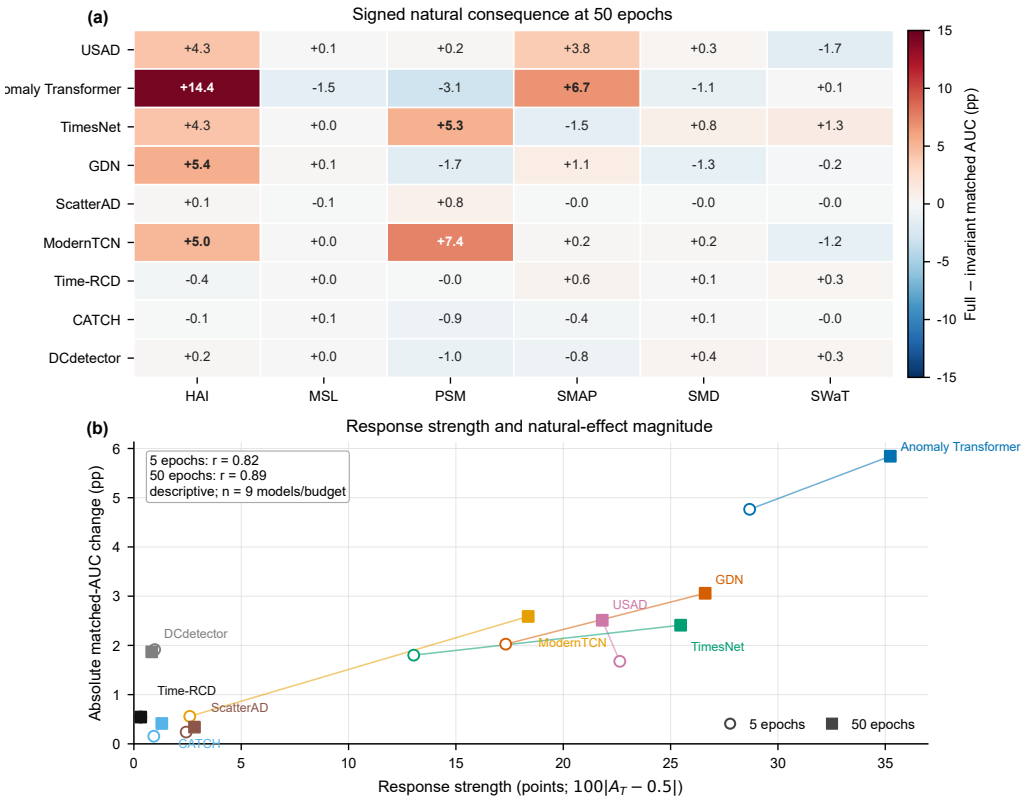

Figure 3: Signed response has dataset-dependent consequences. Top: full-minus-invariant matched-AUC differences by model and dataset. Bottom: response strength versus absolute matched-AUC change, summarized descriptively across nine model means at each budget.

points for Anomaly Transformer, 3.06 for GDN, 2.59 for ModernTCN, and 2.51 for USAD. The association concerns *how much* detection changes; the signed results above show whether the change helps or hurts. Appendix I gives the full model table and paired budget estimates.

5.3 Similar native AUCs conceal large response differences

Similar detection performance, different responses. On the common panel at 5 epochs, Anomaly Transformer and GDN have native AUCs of 0.6874 and 0.6774, only 1.00 point apart. Their response strengths are 0.2868 and 0.1732, a gap of 11.36 points. TimesNet and ModernTCN show the same pattern: native AUCs of 0.6017 and 0.5903 differ by 1.14 points, while responses of 0.1303 and 0.0260 differ by 10.43 points. Capability certification requires nearby benchmark scores to constrain response strengths to be nearby, at specified tolerances. The Anomaly Transformer–GDN pair lies within a two-point native-AUC band but exceeds a ten-point response gap, violating certification at those tolerances. Appendix A gives the full definition and finite-support result.

Supporting evidence from held-out selection. In a parallel response-strength analysis of the same pre-specified 46 folds, we compare two validation rules in the fixed 13-candidate pool: select the highest native AUC (m_N), or the highest response strength (m_C). We then compare their responses on held-out data:

$$\Delta_C = \mathcal{C}_{T,\text{test}}(m_C) - \mathcal{C}_{T,\text{test}}(m_N).$$

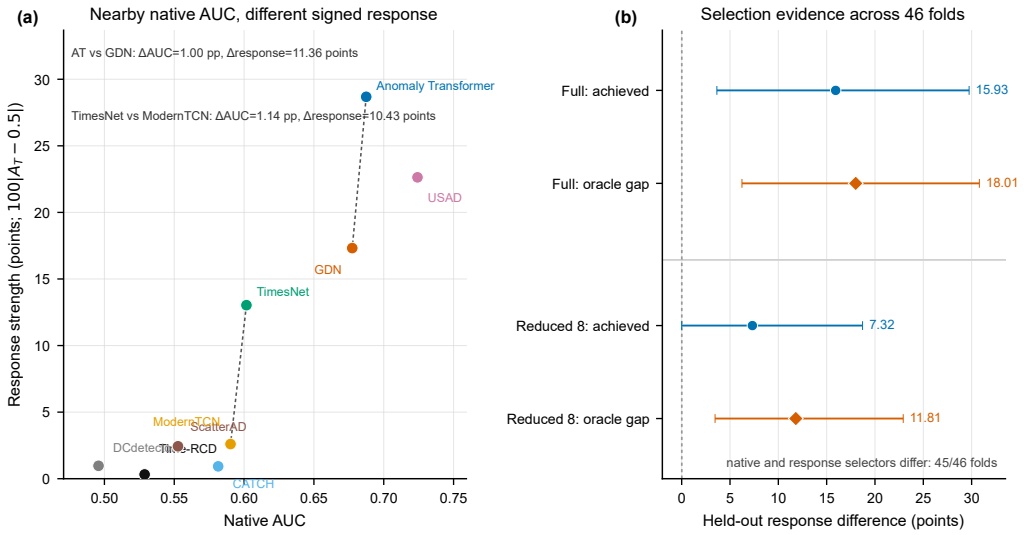

Figure 4: Similar native AUCs leave different signed responses unresolved. Left: nine released models on the common panel. Right: validation-selected response gains and test-oracle gaps in the separate fixed-candidate study, reported as distinct quantities.

Selecting by validation response yields a held-out response that is 15.93 **points higher** on average (95% CI $[3.64, 29.74]$). The result is unchanged at reported precision when we take the worst, mean, or best held-out outcome among numerically tied validation winners. The two validation rules choose different candidates in $45/46$ folds. Unlike the oracle comparison below, this 15.93-point gain is achieved by selecting on validation response.

Distance to the best held-out candidate. We also compare the native-selected candidate with the highest response observed on the test fold:

$$R_C = \max_m \mathcal{C}_{T,\text{test}}(m) - \mathcal{C}_{T,\text{test}}(m_N).$$

This oracle reference gives a mean gap of 18.01 points ($[6.24, 30.80]$). After excluding five scores whose sign behavior is built directly into their definitions, the remaining eight-candidate pool still gives an oracle gap of 11.81 points ($[3.45, 22.92]$). In this subset, response-based selection gives a 7.32-point gain over native-AUC selection, with an interval that includes zero (Appendix F). The pre-specified directional-A_T analysis separately reports 18.01 A_T points of oracle regret ($[6.15, 30.85]$).

6 Discussion

Why can trained scores differ? Several design choices provide plausible routes through which signed cross-channel structure can influence the final score. GDN predicts a sensor from neighboring sensors in a learned graph (Deng & Hooi, 2021). USAD reconstructs multivariate windows (Audibert et al., 2020); changing a joint sign pattern can alter the reconstruction itself and hence the error norm. An error magnitude can therefore remain sensitive to the joint sign pattern through the learned reconstruction. Anomaly Transformer and DCdetector emphasize temporal association discrepancy and dual-attention contrastive learning, respectively (Xu et al., 2022; Yang et al., 2023). These objectives shape the learned mappings, but neither objective directly trains the model on our sign-flip contrast. Under their respective objectives, USAD responds strongly, while DCdetector stays near chance at both budgets. ModernTCN rises from 0.026 to 0.184 with longer training and the architecture fixed (Section 5.1). This budget contrast shows that signed response depends on optimization state and is not fixed by architecture alone.

Responding to the flip does not imply that the response is useful. The natural comparisons hold the trained model, labels, and evaluated positions fixed. Where the full score ranks anomalies better

than its invariant part, the flip-sensitive component helps; where the invariant part ranks them better, that component is misaligned with the labeled dataset. Both outcomes appear across natural datasets, including within the same model family. The selection study compares response strength, whereas the score-removal study measures whether the component helps or harms detection on labeled data.

What native evaluation misses. Native AUC scores the ordering of labeled anomalies and normal examples without separating the contributions of different input distinctions. A detector can rank that mixture well using several available cues, so similar AUCs need not reveal the same signed response. The released-model pairs demonstrate this ambiguity under a shared evaluation protocol (Section 5.3). Evidence that a model detects anomalies well does not, by itself, establish how strongly it responds to this distinction or whether the corresponding component helps on a particular dataset. Reporting A_T and $\mathcal{C}_T$ alongside native AUC exposes the response; comparing s with s_{even} establishes its effect on labeled detection.

Conditions and extension. The shared preprocessing, training-defined channel group, and fixed budgets specify the conditions of the empirical comparison. Focused controls examine marginal and channel-selection effects, and the paired budget study measures change with optimization. Algebraically, the same even–odd split applies to any transformation that restores the input when applied twice. Other transformations would define other response questions; signed orientation is the one evaluated here.

7 Conclusion

Equal magnitude can conceal opposite signed relations. Across nine released TSAD models, the trained scores respond to this distinction with sharply different strengths, and those responses change unevenly with training. The same flip-sensitive score component changes natural anomaly detection, helping on some datasets and harming on others. Similar native AUCs alone do not establish similar signed-orientation responses: the response and its effect on detection require direct measurement.

AI use statement

Generative AI tools were used to assist with writing and language polishing; retrieval and discovery of related work; research ideation and execution, including experimental design, code development, debugging, and result interpretation; and the formulation, checking, and proving of mathematical claims. All AI-assisted outputs were reviewed and revised by the authors. The authors verified the reported experiments, numerical results, citations, and mathematical arguments and take full responsibility for the final content of this submission.

Reproducibility statement

Released-code revisions, preprocessing, channel selection, data eligibility, natural matching, seeds, training budgets, and validation/test splits follow the frozen experiment protocols. The directional held-out study is primary; the parallel $\mathcal{C}_T$, tie-aware, and reduced-pool analyses are post-hoc checks of the same archived outcomes. AUC is computed within each view or segment, and $|A_T - 0.5|$ is formed at seed–view level before averaging. Datasets receive equal weight. Confidence intervals use 50,000 hierarchical-bootstrap replicates, with paired resampling for training-budget comparisons. The appendices give full controls, score definitions, code revisions, dataset support, and secondary estimates.

References

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2485–2494, 2021. doi: 10.1145/3447548.3467174.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3395–3404, 2020. doi: 10.1145/3394486.3403392.

Guilherme O. Campos, Arthur Zimek, Jörg Sander, Ricardo J. G. B. Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E. Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30 (4):891–927, 2016. doi: 10.1007/s10618-015-0444-8.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009. doi: 10.1145/1541880.1541882.

Wenchao Chen, Long Tian, Bo Chen, Liang Dai, Zhibin Duan, and Mingyuan Zhou. Deep variational graph convolutional recurrent network for multivariate time series anomaly detection. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3621–3633, 2022. URL https://proceedings.mlr.press/v162/chen22x.html.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Fereidoon Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021. doi: 10.1609/aaai.v35i5.16523.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.12261.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 387–395, 2018. doi: 10.1145/3219819.3219845.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7194–7201, 2022. doi: 10.1609/aaai.v36i7.20680.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Tian Lan, Hao Duong Le, Jinbo Li, Wenjun He, Meng Wang, Chenghao Liu, and Chen Zhang. Towards foundation models for zero-shot time series anomaly detection: Leveraging synthetic data and relative context discrepancy. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yXqnyIvGAy.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 108231–108261, 2024. doi: 10.52202/079017-3437. Datasets and Benchmarks Track.

Zehao Liu, Mengzhou Gao, and Pengfei Jiao. GCAD: Anomaly detection in multivariate time series from the perspective of granger causality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19041–19049, 2025. doi: 10.1609/aaai.v39i18.34096.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Aditya P. Mathur and Nils Ole Tippenhauer. SWaT: A water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks*, pp. 31–36, 2016. doi: 10.1109/CySWater.2016.7469060.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159, 2020. doi: 10.1145/3368555.3384468.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021. doi: 10.1145/3439950.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron J. Elmore, and Michael J. Franklin. Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(11):2774–2787, 2022. doi: 10.14778/3551793.3551830.

Xiangfei Qiu, Zhe Li, Wanghui Qiu, Shiyan Hu, Lekui Zhou, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Aoying Zhou, Zhenli Sheng, Jilin Hu, Christian S. Jensen, and Bin Yang. TAB: Unified benchmarking of time series anomaly detection methods. *Proceedings of the VLDB Endowment*, 18(9):2775–2789, 2025. doi: 10.14778/3746405.3746407.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022. doi: 10.14778/3538598.3538602.

Hyeok-Ki Shin, Woomyo Lee, Jeong-Han Yun, and HyoungChun Kim. HAI 1.0: HIL-based augmented ICS security dataset. In *13th USENIX Workshop on Cyber Security Experimentation and Test (CSET 20)*. USENIX Association, 2020. URL https://www.usenix.org/conference/cset20/presentation/shin.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2828–2837, 2019. doi: 10.1145/3292500.3330672.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11976.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2b25c39788e5cf11d3541de433ebf4c0-Abstract-Conference.html.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022. URL https://iclr.cc/virtual/2022/spotlight/7024.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3033–3045, 2023. doi: 10.1145/3580305.3599295.

Tao Yin, Shaochen Fu, Zhibin Zhang, Li Huang, Xiaohong Zhang, Yiyuan Yang, Kaixiang Yang, and Meng Yan. ScatterAD: Temporal-topological scattering mechanism for time series anomaly detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1742. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4b52b3c50110fc10f6a1a86055682ea2-Abstract-Conference.html.

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024. doi: 10.1145/3691338.

Xiaona Zhou, Constantin Brif, and Ismini Lourentzou. mTSBench: Benchmarking multivariate time series anomaly detection and model selection at scale. *Transactions on Machine Learning Research*, 2 2026. URL https://openreview.net/forum?id=8LfB8HD1WU.

A FORMAL RESULTS AND CAPABILITY CERTIFICATION

This section gives the formal statements behind the sign-flip reference point and the native-performance comparison.

Theorem 1 (Invariance implies zero response strength)**.** *Suppose $s = h \circ \phi$ and $\phi(T_G x) = \phi(x)$ for all x. Then $A_T(s) = 0.5$ and $\mathcal{C}_T(s) = 0$.*

Proof. Invariance gives $s(T_G x) = h(\phi(T_G x)) = h(\phi(x)) = s(x)$. The original and flipped windows therefore have identical score distributions. Using the Mann–Whitney form of AUC, symmetry assigns equal probability to either strict ordering and half weight to ties, yielding $A_T = 1/2$ (Mann & Whitney, 1947; Hanley & McNeil, 1982). □

Capability certification. The capability in the definition below is the response strength $\mathcal{C}_T$ used throughout the paper.

Definition 1 ((ϵ, δ) capability certification)**.** *Let $\mathcal{B}(s)$ denote a scalar benchmark functional and let $\mathcal{S}$ be a specified class of trained scores. We say that $\mathcal{B}$ (ϵ, δ)-certifies $\mathcal{C}_T$ over $\mathcal{S}$ if, for every $s_1, s_2 \in \mathcal{S}$,*

$$|\mathcal{B}(s_1) - \mathcal{B}(s_2)| \leq \epsilon \quad \Longrightarrow \quad |\mathcal{C}_T(s_1) - \mathcal{C}_T(s_2)| \leq \delta. \tag{7}$$

A violating pair is a witness of non-certification at that tolerance.

Proposition 1 (Finite-score non-identifiability without coverage)**.** *Let a benchmark $\mathcal{B}_D$ depend only on scores at a finite evaluation set D, and let the audit use normal points $U = \{u_i\}_{i=1}^m$ whose transformed points $T_G(U)$ are outside $D \cup U$. If $\mathcal{S}$ is rich enough to contain two score functions that agree on $D \cup U$ but can be assigned independently on $T_G(U)$, then there exist $s_0, s_1 \in \mathcal{S}$ such that*

$$\mathcal{B}_D(s_0) = \mathcal{B}_D(s_1), \qquad \mathcal{C}_T(s_0) = 0, \qquad \mathcal{C}_T(s_1) = \frac{1}{2}.$$

Hence $\mathcal{B}_D$ does not $(0, \delta)$-certify $\mathcal{C}_T$ for any $\delta < 1/2$ over that class.

Proof. Choose s_0 and s_1 to agree on $D \cup U$. Set $s_0(T_G u_i) = s_0(u_i)$ for every i, so the original and transformed audit score multisets are identical and $A_T(s_0) = 1/2$. For s_1, assign every $s_1(T_G u_i)$ above every $s_1(u_j)$ while leaving $D \cup U$ unchanged; then $A_T(s_1) = 1$. The benchmark is unchanged because it sees only D, while the capabilities differ by $1/2$. □

The proposition isolates a source of non-identifiability: the benchmark observes scores on D, while the response contrast also depends on scores at $T_G(U)$. Identifying the response therefore requires either benchmark coverage of the transformed inputs or model-class constraints linking scores on D and $T_G(U)$.

B CONTROLS FOR PREPROCESSING AND THE SIGN-FLIP TEST

The sign-flip test preserves coordinatewise magnitude exactly. Signed marginal symmetry is checked empirically. The focused controls below test alternative explanations for the measured response; they are separate from the full released-model panel.

Rank-Gaussian calibration and independent marginals. All residual channels in the primary analysis are calibrated from the training empirical CDF to a rank-Gaussian scale before G is chosen or the sign flip is applied. A channelwise empirical-CDF scorer that uses independent marginal information remains near chance in the focused sanity panel. We also audit pre/post-transform channel means, standard deviations, and positive rates to quantify signed marginal behavior directly.

Exact sign-invariant scores. On matched original/transformed windows in the focused sanity panel, Fisher, VAR-L2, Abs-Copula, and Branch-Mag are identical up to numerical precision and obtain audit AUC 0.500 ± 0.000. This is the empirical counterpart of Theorem 1 and rules out threshold/tie handling as an explanation for their chance result.

Coordinate-selection and relational-shuffle controls. The primary group is the training-only top-correlation group with $|G| = \min(8, d-1)$. Random and low-correlation group ablations retain the qualitative signed-vs-invariant separation across alternative coordinate selections. A separate relational-shuffle control preserves channelwise empirical values while breaking cross-channel alignment; it is used as a falsification test for a purely independent-marginal explanation.

C RELEASED-MODEL PROVENANCE

Table 3: Released-code revisions used in the sign-flip study.

Model	Frozen commit
Anomaly Transformer	`b0ee470c8012bff857fb600462aec6209c4a18d9`
TimesNet / Time-Series-Library	`4e938a1767106324dd753b2a44832bf870a0252e`
USAD	`e25af45c8e1c32783aed94fc7e5ab85effef2b1a`
GDN	`9853899da860682669a134e4af315d036aab4eca`
ModernTCN	`56a9a2c018385cd5acef015378cae7f084d1b11c`
DCdetector	`9d107cae518ef24ee2eb2ad1b92107cdcd09ea10`
CATCH	`3647c69be5eb56649b072596cf89098e689e20c3`
ScatterAD	`7572fa954468e46c598badf1c9cfaf6c1835c4cb`
Time-RCD	`372bb980426b2f67007311c6f3165ab789c79bef`

OmniAnomaly was excluded before outcomes because its legacy dependencies were incompatible with the pre-specified software environment; no replacement was added after outcomes were observed. Time-RCD is evaluated in its released zero-shot form. The controlled response panel retains each released model's native architecture, objective, and anomaly score using the same preprocessed inputs.

D CONTROLLED 17-SCORE PORTFOLIO

To examine response differences among scores with similar native AUC, we use the following 17 scoring rules, fixed before computing the comparison:

Score	Score
Abs-Copula	Branch-Corr
Branch-Local	Branch-Mag
Branch-Signed	CorrShift-Fro
Fisher	IsolationForest
KNN	LOF
Portfolio-AuditGated	Portfolio-Max
Portfolio-Mean	Portfolio-Top2Mean
Signed-Copula	Signed-Mahal
VAR-L2	—

The composite rules are fixed arithmetic combinations of train-referenced percentile scores on a common $[0, 1]$ scale. `Branch-Mag` averages Fisher, VAR-L2, and Abs-Copula; `Branch-Signed` averages Signed-Mahal, Signed-Copula, and Signed-CrossCov; `Branch-Corr` averages CorrShift-Fro and FisherZ-CorrShift; and `Branch-Local` averages KNN, LOF, and IsolationForest. `Portfolio-Mean` averages the four branch scores, `Portfolio-Max` takes their maximum, and `Portfolio-Top2Mean` averages the two largest branches. Signed-CrossCov and FisherZ-CorrShift are internal branch components of these rules. `Portfolio-AuditGated` uses only the current branch scores: it applies the balanced branch mean when $\max(\text{Branch-Signed}, \text{Branch-Corr}) < 0.90$ and otherwise applies fixed Signed/Corr/Local/Mag weights $0.40/0.35/0.15/0.10$. This controlled rule stress-tests mechanism-sensitive composition; the released-model evidence is established independently on the nine-model panel. The native-AUC comparison contains 64 views across MSL, SMD, SMAP, and SWaT; the response-diameter calculation uses the shared native/response support rather than the additional response-only view.

Table 4: Native-equivalence diameter in the 17-score portfolio. Response strength is $|\mathrm{AUC}_T - 0.5|$; recall diameters use the directional-response operating point.

ϵ	Candidates	Native AUC range	Response diameter	Recall@5% diameter
0.005	2	[0.6802, 0.6814]	0.0079	0.1104
0.010	2	[0.6802, 0.6814]	0.0079	0.1104
0.020	5	[0.6635, 0.6814]	0.0882	0.1291

E SELECTION MODEL POOL AND AGGREGATION

Table 5: Thirteen pre-specified candidate families used in the held-out selection study.

Candidate	Family	Training
Fisher	magnitude/statistical	deterministic
VAR-L2	forecast-residual magnitude	deterministic
Abs-Copula	sign-erasing dependence	deterministic
Signed-Mahal	signed covariance	deterministic
Signed-Copula	signed rank dependence	deterministic
CorrShift-Fro	correlation shift	deterministic
PCA-Stats	linear reconstruction	deterministic
KNN-Stats	local distance	deterministic
IF-Stats	isolation/tree	stochastic
OCSVM-Stats	kernel one-class	deterministic
AE-Seq	deep reconstruction	stochastic
GRU-Predict	deep temporal forecasting	stochastic
Transformer-Predict	attention forecasting	stochastic

Stochastic models use seeds $0, 1, 2$. Score orientation was specified in advance so that higher values always indicate greater anomaly. We use the same hyperparameters across views and compute metrics separately within each view. Directional analyses aggregate A_T according to the pre-specified protocol; response-strength analyses first form $\mathcal{C}_T = |A_T - 0.5|$ for each seed–view pair and only then average across seeds and higher-level units. For HAI selection folds, views are averaged within each release block and release-block means are then weighted equally; MSL/SMD/SMAP use equal views, while SWaT/PSM use their pre-specified event-time segments. Cross-domain inference treats domain as the top-level unit; views, events, matched pairs, and overlapping windows remain nested evidence units.

F RESPONSE-STRENGTH SELECTION ROBUSTNESS

The primary held-out analysis was pre-specified using directional audit AUC A_T. Because the certification definition in Appendix A uses response strength $\mathcal{C}_T = |A_T - 0.5|$, we ran a parallel post-hoc robustness analysis on the same 46 folds and the same archived candidate outcomes. No candidate was retrained and no fold, seed, native metric, or candidate pool was changed. The post-hoc $\mathcal{C}_T$ analysis forms $|A_T - 0.5|$ for each seed–view pair before averaging the archived outcomes.

The native selector remains

$$m_N = \arg\max_m \mathrm{AUC}_{\mathrm{native,val}}(m),$$

while the response-strength selector is

$$m_C = \arg\max_m \mathcal{C}_{T,\mathrm{val}}(m).$$

We report

$$R_C = \max_m \mathcal{C}_{T,\mathrm{test}}(m) - \mathcal{C}_{T,\mathrm{test}}(m_N),$$

the response-strength oracle regret of native selection, and

$$\Delta_C = \mathcal{C}_{T,\mathrm{test}}(m_C) - \mathcal{C}_{T,\mathrm{test}}(m_N),$$

the held-out contrast between response-based and native-AUC validation selection.

Full candidate pool. Across the 46 folds, the native and $\mathcal{C}_T$ validation selectors differ in $45/46$ folds. Averaging datasets equally, the model selected by native AUC falls 18.01 response points below the best candidate on the held-out test folds (95% hierarchical-bootstrap CI $[6.24, 30.80]$). The fold-pooled median regret is 20.46 points and the maximum is 50.00 points; regret is positive in $40/46$ folds, exceeds 5 points in $28/46$, 10 points in $26/46$, and 20 points in $23/46$.

Tie-aware validation selection. Seven folds contain more than one numerically tied validation maximizer of $\mathcal{C}_T$. Ties use absolute tolerance 10^{-4} and zero relative tolerance. We therefore evaluate worst, mean, and best held-out Δ_C over all tied maximizers. After averaging datasets equally, the worst, mean, and best tie-handling rules all give a 15.93-point contrast at reported precision, with 95% hierarchical-bootstrap intervals $[3.64, 29.74]$ in each case. Thus the full-pool response-selection conclusion does not depend on an arbitrary numerical tie rule.

Reduced detector-oriented subset. We remove Fisher, VAR-L2, Abs-Copula, Signed-Mahal, and Signed-Copula because their relation to the audit is built directly into their score definitions. The remaining eight candidates are CorrShift-Fro, PCA-Stats, KNN-Stats, IF-Stats, OCSVM-Stats, AE-Seq, GRU-Predict, and Transformer-Predict. After averaging datasets equally, native selection remains 11.81 response points below the test oracle, with 95% CI $[3.45, 22.92]$. The corresponding validation-$\mathcal{C}_T$ selection contrast is 7.32 points; its tie-aware 95% interval extends from slightly below zero (rounding to -0.00) to 18.72 points under all three rules at reported precision. The reduced-subset evidence therefore establishes a response-strength shortfall relative to the test oracle; the selection-contrast interval includes zero.

Table 6: Parallel $\mathcal{C}_T$ held-out robustness analysis on the pre-specified 46 folds. Values are response points; intervals are equal-dataset hierarchical-bootstrap 95% CIs.

Pool	Folds	Selector change	Mean R_C	95% CI
Full 13-family pool	46	45/46	18.01	[6.24, 30.80]
Reduced 8-candidate subset	46	45/46	11.81	[3.45, 22.92]

Table 7: Tie-aware held-out response-selection contrast Δ_C in the full pool. Worst/mean/best refer to test performance over all numerically tied validation-$\mathcal{C}_T$ maximizers within each fold.

Tie handling	Mean Δ_C	95% CI
Worst	15.93	[3.64, 29.74]
Mean	15.93	[3.64, 29.74]
Best	15.93	[3.64, 29.74]

G NATURAL-DOMAIN SUPPORT

Table 8: Datasets used to compare the full score with its flip-invariant part. Counts are views/files, not independent cross-domain replicates.

Domain	Views	Source
HAI	13	HAI benchmark (Shin et al., 2020)
MSL	7	NASA telemetry (Hundman et al., 2018)
PSM	1	Pooled Server Metrics released with RANSynCoders (Abdulaal et al., 2021)
SMAP	13	NASA telemetry (Hundman et al., 2018)
SMD	21	server-machine data released with OmniAnomaly (Su et al., 2019)
SWaT	1	Secure Water Treatment testbed (Mathur & Tippenhauer, 2016)
Total	56	

H RELEASED-MODEL NATIVE-EQUIVALENCE WITNESSES ON THE COMMON PANEL

Table 9: Two direct released-model witnesses on the common 56-view, 5-epoch panel. Values are equal-dataset summaries from the same model runs and show the response-strength gap within the native-AUC equivalence band.

Model	Native AUC	$\mathcal{C}_T$
Anomaly Transformer	0.6874	0.2868
GDN	0.6774	0.1732
TimesNet	0.6017	0.1303
ModernTCN	0.5903	0.0260

I NATURAL DETECTION CHANGES AND TRAINING-BUDGET SUMMARIES

Table 10: Natural consequence of removing the flip-sensitive score component in the pre-specified high-response subgroup. Entries are absolute percentage-point changes; the final column is the paired 50-minus-5 contrast.

Metric	5 epochs	50 epochs	Paired $50-5$
Native AUC	1.75 [1.04, 2.96]	2.63 [1.63, 4.32]	+0.88 [0.11, 1.84]
Matched AUC	2.57 [1.74, 3.72]	3.46 [1.87, 5.48]	+0.89 [−0.17, 2.10]
Unmatched window recall@5%	2.57 [1.64, 3.68]	4.11 [2.12, 6.68]	+1.55 [−0.17, 3.53]
Matched window recall@5%	2.56 [1.47, 4.13]	3.45 [2.11, 5.12]	+0.89 [−0.31, 2.00]
Event recall@5%	5.26 [2.72, 8.96]	5.32 [3.03, 8.06]	+0.06 [−3.19, 3.27]

Table 11: Response strength and mean absolute magnitude-matched AUC change on the common 56-view panel. Datasets receive equal weight. Time-RCD is zero-shot and shared across budgets.

Model	$\mathcal{C}_T$ 5ep	$\mathcal{C}_T$ 50ep	$\lvert D_{\mathrm{matched}}\rvert$ 5ep	$\lvert D_{\mathrm{matched}}\rvert$ 50ep
Anomaly Transformer	0.2868	0.3523	0.0476	0.0584
GDN	0.1732	0.2661	0.0203	0.0306
TimesNet	0.1303	0.2547	0.0180	0.0241
USAD	0.2263	0.2182	0.0168	0.0251
ModernTCN	0.0260	0.1837	0.0056	0.0259
DCdetector	0.0097	0.0083	0.0192	0.0187
CATCH	0.0093	0.0130	0.0015	0.0041
ScatterAD	0.0244	0.0282	0.0024	0.0034
Time-RCD	0.0032	0.0032	0.0054	0.0054

Across these nine model points, the descriptive Pearson association between response strength and absolute matched-AUC change is approximately 0.82 at 5 epochs and 0.89 at 50 epochs.

J DIRECTIONAL AUDIT SUMMARIES

The directional audit statistic is A_T (called *pure-sign AUC* in the frozen experiment logs), while the paper's headline response strength is $\mathcal{C}_T = |A_T - 0.5|$. The absolute deviation records response strength, while A_T retains response direction. The underlying directional audit remains informative: models differ in the frequency with which individual common-support views are near chance, positively sensitive, negatively sensitive, or mixed across seeds. For example, on the strict 65-view common support, DCdetector is near chance on 64 views and CATCH on 63, whereas Anomaly Transformer is near chance on 14 and positively sensitive on 34; the remaining views can be mixed or negatively sensitive. These view-level counts expose directional heterogeneity compressed by the equal-dataset mean.

K ESTIMANDS AND INTERPRETATION

Controlled response strength. $\mathcal{C}_T(s) = |\mathrm{AUC}_T(s) - 0.5|$ measures response strength under the specified sign flip, irrespective of direction.

Response range at similar native AUC. Within an ϵ band of native AUC, the range of $\mathcal{C}_T$ values shows how much signed response remains unresolved by native performance for this model class.

Directional and response-strength selection diagnostics. The pre-specified 46-fold study uses directional audit AUC A_T directly: m_S maximizes validation A_T, and its oracle regret is $\max_m A_{T,\mathrm{test}}(m) - A_{T,\mathrm{test}}(m_N)$. A parallel post-hoc robustness analysis instead evaluates $\mathcal{C}_T = |A_T - 0.5|$ on the same archived folds; Appendix F reports response-strength oracle regret, tie-aware validation selection, and the reduced eight-candidate subset. The two analyses therefore evaluate response direction and response magnitude separately. In a separate analysis across six datasets, we control for magnitude and evaluate detection on the original labeled data. The AUC change ranges from -12.31 to $+9.36$ points across datasets, with a mean absolute change of 5.46 points. This secondary result is reported with the selection diagnostic; the main natural analysis uses the released-model 56-view panel.

Natural signed consequence. $D_{\mathrm{nat}} = \mathrm{AUC}(y, s) - \mathrm{AUC}(y, s_{\mathrm{even}})$ is the full-minus-invariant AUC difference for the same trained score on a fixed labeled dataset. A positive value means that the flip-sensitive component improves AUC; a negative value means that it reduces AUC. Its absolute value measures the size of the effect.

Optimization contrast. The 50-minus-5 comparison is paired on model, domain, view, and seed. It estimates sensitivity to this specific optimization budget change at the two evaluated budgets.

Fixed-FPR metrics. Recall comparisons use thresholds estimated from unmatched evaluation normals, or from the corresponding matched normal controls for magnitude-matched data, at the same empirical operating FPR for each score being evaluated. These recall values describe performance at the specified empirical FPR. Event recall uses pre-specified contiguous attack-window clusters as the event unit.

L CHANNEL-GROUP SENSITIVITY ON THE RELEASED-MODEL PANEL

We vary the channel group on the 65-view, five-dataset panel using the original 5-epoch released-model protocol. The eight settings are top-correlation groups with requested sizes $k \in \{1, 2, 4, 8\}$, a low-correlation group with $k = 8$, and three random-group replicates with $k = 8$. The effective size is $\min(k, d - 1)$ in each view. The pre-specified top- and low-correlation selectors rank channels by the sum of absolute training correlations; random groups sample channels uniformly without replacement. Each random group is shared across models and training seeds within a view, with a separate group seed for each replicate.

Preprocessing, training inputs, model implementations, scoring rules, and sampling limits follow the primary response study. For each setting, the eight trainable models were rerun for 5 epochs with seeds $0, 1, 2$ and batch size 64; Time-RCD uses the same released zero-shot checkpoint, seed 0, and 256-pair cap. We form $\mathcal{C}_T = |A_T - 0.5|$ for each seed and view before averaging over seeds, views within datasets, and datasets with equal weight. Group seeds follow $101003 + 10007i + 1000003r$, where i is the original file-list index and $r \in \{0, 1, 2\}$ is the random-group replicate.

Table 12 summarizes each setting, and Table 13 reports all nine model means. Across the seven alternatives, rank correlations with the rerun canonical setting range from 0.9167 to 0.9833. USAD, Anomaly Transformer, TimesNet, and GDN occupy the top four positions in every setting, although their order and absolute responses change. The random-group mean responses are 0.1067, 0.1024, and 0.1047, compared with 0.0830 under the canonical group. The response differences therefore also occur without correlation-based channel selection. These are descriptive comparisons of the fixed nine-model panel.

Table 12: Channel-group sensitivity at 5 epochs on 65 views from five datasets. k is the requested size, capped at $d - 1$. Rank correlations compare the nine model means with the rerun top-correlation $k = 8$ setting. Mean and spread refer to response strength on its $[0, 0.5]$ scale; spread is the maximum minus minimum across models.

Selection	k	Replicate	Spearman ρ	Mean	Spread
Top-correlation	1	—	0.9167	0.0520	0.1330
Top-correlation	2	—	0.9167	0.0712	0.1892
Top-correlation	4	—	0.9333	0.0842	0.2038
Top-correlation	8	—	1.0000	0.0830	0.2222
Low-correlation	8	—	0.9667	0.0716	0.2082
Random	8	0	0.9500	0.1067	0.2963
Random	8	1	0.9833	0.1024	0.2775
Random	8	2	0.9667	0.1047	0.3041

Table 13: Response strength for each model and channel-group setting. Top-k and Low-8 use training-correlation selection; R0–R2 are random-group replicates at requested $k = 8$. Values use the same seed–view–dataset aggregation as the primary response study.

Model	Top-1	Top-2	Top-4	Top-8	Low-8	R0	R1	R2
USAD	0.1246	0.1646	0.2159	0.2333	0.2192	0.2295	0.2490	0.2400
Anomaly Transformer	0.1393	0.1983	0.2060	0.1740	0.1514	0.3134	0.2908	0.3199
TimesNet	0.0778	0.1165	0.1348	0.1208	0.0684	0.1608	0.1474	0.1536
GDN	0.0623	0.0753	0.0984	0.1171	0.1253	0.1253	0.1178	0.1053
ScatterAD	0.0232	0.0283	0.0287	0.0299	0.0237	0.0348	0.0365	0.0277
ModernTCN	0.0152	0.0256	0.0338	0.0263	0.0142	0.0401	0.0323	0.0375
Time-RCD	0.0063	0.0091	0.0121	0.0202	0.0202	0.0220	0.0190	0.0255
CATCH	0.0066	0.0095	0.0146	0.0144	0.0114	0.0171	0.0159	0.0169
DCdetector	0.0128	0.0136	0.0137	0.0111	0.0110	0.0173	0.0133	0.0158