DEPENDENCE-AWARE MULTI-HEAD ACTIVATION MONITORING FOR DISTRIBUTION SHIFT AND OOD RELIABILITY

Anonymous authors
Paper under review

ABSTRACT

Modern neural systems frequently fail under deployment shift because confidence-only diagnostics underrepresent hidden changes in internal activations, and static benchmark metrics do not directly control sequential alert burden. We study a reusable, architecture-agnostic activation-monitoring framework that combines three detector families at each layer: confidence-energy scores, distance-based geometry scores, and streaming discrepancy scores. The method introduces dependence-aware calibration, where global alert thresholds are inflated by online variance and autocorrelation estimates, and a boundary-aware ranking analysis that clarifies when energy-based and discrepancy-based detector orderings agree or diverge. Across static OOD and streaming stress evaluations, fused monitoring improves near-OOD AUROC by 0.013 and far-OOD AUROC by 0.020 over the strongest single-head energy comparator while preserving matched-recall false-alert parity in this evaluation. Time-to-detection improves from 10.87 to 5.40 windows relative to the energy baseline, and boundary-aware selection reduces policy regret by 21.0% on average over five seeds. Symbolic validation verifies 11 of 12 theorem obligations; the kernel-regime inversion result is supported for $\gamma \in \{1, 2, 4\}$ and fails at $\gamma = 0.5$, which explicitly bounds generality. These results support a conditional conclusion: multi-head, dependence-aware activation monitoring improves detection utility and operational interpretability, while full generality still depends on unresolved low-bandwidth and real-data replay regimes.

1 INTRODUCTION

Silent degradation under distribution shift is one of the main blockers to dependable deployment of modern machine learning systems. The failure mode is particularly severe when score distributions drift gradually: the model remains confident enough to avoid trivial alarms, yet internal representations move away from calibration conditions established during model development. This failure pattern is visible across vision, language, and domain-specialized workloads and is repeatedly observed in both OOD and drift literatures (Anthony et al., 2026; Ekim et al., 2024; Song et al., 2023; Ltd., 2020; Rabanser et al., 2019). A practical monitoring layer therefore needs to do more than compute one confidence statistic; it must integrate diverse signals, preserve low intervention overhead, and expose interpretable failure boundaries.

Recent OOD work offers strong detector families, including confidence and temperature-scaling methods (Hendrycks & Gimpel, 2017; Liang et al., 2018; Hsu et al., 2020), energy formulations (Liu et al., 2020; Wu et al., 2023; Guglielmo & Masana, 2025), distance-geometry methods (Lee et al., 2018; Anthony & Kamnitsas, 2023; Yahiaoui et al., 2026), activation-structure methods (Sastry & Oore, 2019; Zongur et al., 2025; Wan et al., 2024), and representation-shaping approaches (Harun et al., 2025; Wu et al., 2024; Haas et al., 2022). Drift detection work provides complementary sequential tests and operational toolchains (Ayers et al., 2025; Varshney & Torra, 2024; Ltd., 2020). The central unresolved issue is not detector availability but claim transport: methods that rank well on static benchmarks can fail to maintain stable alert burden under autocorrelated or contaminated streams (Anthony et al., 2026; Ekim et al., 2024; Rabanser et al., 2019). This discrepancy motivates a unified monitoring perspective where static discrimination and sequential reliability are co-optimized instead of treated as separate tasks.

This paper develops and evaluates such a perspective for activation monitoring. We build a multi-head framework that computes per-layer scores from three families (energy/confidence, Mahalanobis-style geometry, and discrepancy testing), calibrates each head on reference windows, and aggregates them through a dependence-aware global threshold. The framework is designed as a reusable instrumentation layer rather than a model-specific retraining pipeline, so it aligns with common deployment constraints while preserving gradient-path safety and architecture-agnostic hook compatibility.

Our study is deliberately hybrid: it includes formal analysis to clarify assumptions and operating boundaries, and it includes executed simulation evidence to measure practical gains and expose unsupported regions. The formal component establishes monotonicity and perturbation properties of autocorrelation-adjusted thresholds, plus conditional equivalence and non-equivalence boundaries between energy and discrepancy rankers. The empirical component tests the same claims under static OOD and streaming stress with explicit seed-level uncertainty.

The work matters beyond OOD benchmarking. In safety-critical pipelines, operators need monitors that explain why alerts change when data dependence changes. In large-scale experimentation, researchers need detector comparisons that remain meaningful when latency, false-alert budgets, and window dependence interact. In model-agnostic observability, developers need monitoring modules that can be attached to CNN, MLP, and Transformer components without architecture-specific redesign.

Our main contributions are:

- We introduce a dependence-aware multi-head activation monitoring objective that jointly optimizes static OOD separation and sequential alert quality while keeping calibration and decision variables explicit.
- We provide theorem-backed boundary analysis showing when energy and discrepancy ranking policies are conditionally equivalent and when kernel-regime effects produce constructive non-equivalence.
- We execute a unified validation protocol across static and streaming regimes and report a mixed but informative evidence profile: improved AUROC and time-to-detection, policy-regret reduction, and explicit caveats where strict FAR superiority is not yet achieved.
- We publish an auditable claim-to-evidence structure that links each major conclusion to figures, tables, symbolic checks, and unresolved gaps, enabling targeted reruns rather than narrative overclaiming.

2 Related Work and Gap Synthesis

2.1 Confidence and Energy Lineages

Confidence-threshold monitoring remains the operational baseline for many systems because of simplicity and low overhead. Maximum softmax probability (MSP) provides a reference anchor (Hendrycks & Gimpel, 2017); ODIN-style perturbation and temperature variants improve separability in many settings (Liang et al., 2018; Hsu et al., 2020); and energy-based scoring generalizes confidence behavior through a log-sum-exp potential that is often more stable under near-OOD shifts (Liu et al., 2020; Wu et al., 2023; Guglielmo & Masana, 2025). Additional calibration variants, including virtual-logit matching and outlier exposure controls, further improve score shaping (Sun et al., 2022; Papadopoulos et al., 2019; Wu et al., 2022).

Despite these advances, confidence-energy families remain sensitive to calibration-window representativeness and can miss hidden-bias shifts where score magnitudes remain deceptively well-calibrated (Anthony et al., 2026; Ekim et al., 2024; Rabanser et al., 2019). This sensitivity appears repeatedly in high-confidence failure analyses and near-OOD adversarial settings (Fort, 2022; Uwimana1 & Senanayake, 2021; Chen et al., 2020). The implication for activation monitoring is not that confidence heads are obsolete, but that they should be one component in a broader score family.

2.2 Geometry, Activation Structure, and Representation Effects

Feature-space methods use richer representation information than confidence-only scores. Mahalanobis detectors and related unified formulations convert class-conditional geometry into anomaly scores and often provide stronger near-OOD behavior when assumptions hold (Lee et al., 2018; Anthony & Kamnitsas, 2023; Yahiaoui et al., 2026). Activation-structure approaches such as Gram statistics, activation priors, channelwise aggregation, and subspace methods exploit intermediate features directly (Sastry & Oore, 2019; Wan et al., 2024; Tuononen et al., 2025; Zongur et al., 2025). ReAct and related clipping methods show that modest activation interventions can improve OOD detection with low implementation complexity (Sun et al., 2021; Guglielmo & Masana, 2025).

A second cluster links representation quality to OOD reliability via neural-collapse or separation viewpoints (Harun et al., 2025; Wu et al., 2024; Ammar et al., 2023; Liu & Qin, 2023; Haas et al., 2022; Wang et al., 2025). These methods motivate deeper integration between monitoring and representation geometry, but they also increase assumption burden and may require training-time intervention. For a reusable monitoring layer, this creates a tradeoff: stronger separability potential versus wider operational constraints.

2.3 Streaming Drift Detection and Operational Reliability

Drift-detection literature emphasizes sequential error control, alert latency, and nonstationarity management (Ayers et al., 2025; Varshney & Torra, 2024; Ltd., 2020). Shift studies show that offline metrics alone can be insufficient for deployment reliability, especially when dependence, seasonality, and window contamination alter false-alert behavior (Rabanser et al., 2019; Ekim et al., 2024). Tooling ecosystems such as OpenOOD and alibi-detect make these evaluations reproducible and practical (Yang et al., 2023; Ltd., 2020).

However, there is still a gap between OOD and drift paradigms. OOD studies often optimize AUROC/FPR95 under fixed splits, while drift studies optimize detection delay and alarm rates under sequential windows. The combined literature identifies this mismatch as a blocking issue for scientific claims that aim to describe deployment behavior. A monitoring architecture that cannot reconcile both metrics risks reporting improvements that are not actionable in streams.

2.4 Boundary of Novelty

Our contribution sits at the boundary between established detector primitives and unresolved cross-family interoperability. The reused primitives include energy and confidence scores (Liu et al., 2020; Hendrycks & Gimpel, 2017), Mahalanobis geometry (Lee et al., 2018; Anthony & Kamnitsas, 2023), and discrepancy testing (Rabanser et al., 2019; Ayers et al., 2025). The new contribution is a joint formulation that (i) calibrates and aggregates these heads with dependence-aware thresholds, and (ii) makes equivalence conditions explicit rather than implicit when comparing detector families. This is distinct from proposing a new score in isolation; instead, we provide an auditable framework for when score families should or should not be expected to agree.

3 Problem Setting and Assumptions

3.1 Notation and Borrowed Scoring Primitives

Let x_t denote an input (or batch summary) at stream index $t \in \{1, \dots, T\}$, and let $\mathcal{L}$ be the set of monitored layers. For each $l \in \mathcal{L}$ we observe activation features $f_l(x_t) \in \mathbb{R}^{d_l}$. We reuse established detector primitives at each layer.

First, the energy score for logits $z(x_t) \in \mathbb{R}^C$ is

$$E_t = -\tau \log \sum_{c=1}^{C} \exp\left(\frac{z_c(x_t)}{\tau}\right), \tag{1}$$

which follows the standard energy formulation (Liu et al., 2020; Wu et al., 2023). Second, class-conditional Mahalanobis scoring is defined by

$$M_t = \max_c \left[-\left(f_l(x_t) - \mu_{l,c}\right)^\top \Sigma_l^{-1} \left(f_l(x_t) - \mu_{l,c}\right) \right], \tag{2}$$

using the common class-mean and covariance structure (Lee et al., 2018; Anthony & Kamnitsas, 2023). Third, we use window discrepancy statistics Q_t versus reference Q_{ref} via MMD-type tests,

$$\text{MMD}^2(Q_t, Q_{\text{ref}}) = \left\| \mathbb{E}_{Q_t}[\phi(x)] - \mathbb{E}_{Q_{\text{ref}}}[\phi(x)] \right\|_{\mathcal{H}}^2, \tag{3}$$

as in two-sample drift testing (Rabanser et al., 2019; Ayers et al., 2025).

3.2 Manuscript-Specific Monitoring Variables

In this work we define calibrated per-head layer scores

$$\widetilde{S}_{l,t}^h = \frac{S_{l,t}^h - \widehat{\mu}_{l,h}}{\widehat{\sigma}_{l,h} + \epsilon}, \quad h \in \{E, M, Q\}, \tag{4}$$

where $(\widehat{\mu}_{l,h}, \widehat{\sigma}_{l,h})$ are warmup estimates and $\epsilon > 0$ avoids division by zero. The layer-aggregated score and global score are

$$\widetilde{S}_{l,t} = \sum_h w_{l,h} \widetilde{S}_{l,t}^h, \qquad G_t = \sum_{l \in \mathcal{L}} \alpha_l \widetilde{S}_{l,t}, \tag{5}$$

with simplex constraints $w_{l,h} \geq 0$, $\sum_h w_{l,h} = 1$, $\alpha_l \geq 0$, $\sum_l \alpha_l = 1$.

To account for dependence, we define an adaptive threshold

$$\tau_t = z_{1-\alpha}\sqrt{\widehat{v}_t}, \qquad \widehat{v}_t = \widehat{\sigma}_t^2\left(1 + 2\sum_{k=1}^{K}\widehat{\rho}_{t,k}\right), \tag{6}$$

where $\widehat{\rho}_{t,k}$ is lag-k autocorrelation and α is target tail probability. An alert is raised when $G_t > \tau_t$ after warmup and cooldown gates are satisfied.

3.3 Objective, Feasible Set, and Optimality Criterion

We formalize monitoring as constrained optimization over decision variables

$$\theta = \{w_{l,h}, \alpha_l, \alpha, K, W, \text{cooldown}\}, \tag{7}$$

where W is the scoring window size. Let $\mathrm{Rec}(\theta)$ denote shift recall, $\mathrm{FAR}(\theta)$ false alerts per 1k windows, $\mathrm{TTD}(\theta)$ time to detect, and $\mathrm{OH}(\theta)$ normalized compute-memory overhead. We minimize

$$J(\theta) = -\lambda_1\mathrm{Rec}(\theta) + \lambda_2\mathrm{FAR}(\theta) + \lambda_3\mathrm{TTD}(\theta) + \lambda_4\mathrm{OH}(\theta), \tag{8}$$

subject to

$$\Theta = \{\theta : \mathrm{FAR}(\theta) \le B_{\mathrm{far}},\ \mathrm{OH}(\theta) \le B_{\mathrm{oh}},\ w, \alpha \text{ satisfy simplex constraints}\}. \tag{9}$$

A monitor is optimal in this paper if it is Pareto-efficient in $(\mathrm{Rec}, \mathrm{FAR}, \mathrm{TTD}, \mathrm{OH})$ within Θ and minimizes equation 8 under the selected weight vector (λ_i). In the executed run, overhead constraints are treated as deployment metadata and verified separately in reproducibility records.

4 Method: Dependence-Aware Multi-Head Activation Monitoring

4.1 Calibration Pipeline

The method uses a warmup phase to estimate head-wise moments and dependence statistics before alerting is activated. Each layer-head channel stores running first and second moments and, for discrepancy channels, reference-window embeddings required by equation 3. Normalization via equation 4 makes head outputs commensurate even when native scales differ by orders of magnitude. This is crucial when combining confidence and geometry terms; without normalization, one head can dominate aggregation purely by scale.

Head weights $w_{l,h}$ can be fixed or selected through constrained search on calibration windows. In this run we evaluate both single-head and fused variants, and we keep the fused variant transparent by enforcing nonnegative simplex weights. The simplex restriction improves interpretability: each alert score can be decomposed into convex contributions from known detector families.

4.2 Dependence-Aware Global Decision Rule

Static thresholds assume independent windows and can underestimate tail probability under autocorrelation. The threshold in equation 6 inflates nominal variance by a Newey-West-like dependence factor. The resulting decision rule is conservative under high dependence and converges to the independent-window threshold when autocorrelation terms vanish. This behavior is central to our operational claim because false-alert budgets are measured sequentially, not i.i.d.

Warmup contamination can still bias estimates; therefore the pipeline couples threshold adaptation with contamination diagnostics. In stress sweeps we track residuals between observed and expected alert rates and flag regimes where assumptions are violated. These flags are not treated as post hoc decorations; they are part of the method output and directly gate interpretation of alert-rate claims.

4.3 Boundary-Aware Detector Ranking

A practical monitor often needs a policy for selecting among candidate score heads or combinations. Let Δe_t be an energy-shift statistic and D_t a discrepancy statistic. Under assumptions described in section 5, if

$$D_t = a\Delta e_t + b + \xi_t, \qquad a > 0, \quad \mathbb{E}[\xi_t \mid \Delta e_t] = 0, \tag{10}$$

then ranking by D_t and ranking by Δe_t are equivalent in expectation. This equivalence provides a criterion for when detector substitution is justified. Conversely, kernelized discrepancy can violate this affine bridge in finite windows, motivating explicit non-equivalence diagnostics.

Algorithm 1 Dependence-aware multi-head activation monitoring

1: **Input:** monitored layers $\mathcal{L}$, head set $\{E, M, Q\}$, warmup length W_0, window size W, lag cap K, alert level α.
2: Collect warmup activations and compute head scores per layer.
3: Estimate $(\widehat{\mu}_{l,h}, \widehat{\sigma}_{l,h})$ and reference discrepancy statistics for each (l, h).
4: For each stream index t: compute normalized head scores using equation 4.
5: Aggregate head scores to layer scores, then to global score G_t using equation 5.
6: Update dependence estimates $(\widehat{\sigma}_t^2, \widehat{\rho}_{t,1:K})$ and threshold τ_t by equation 6.
7: If cooldown is inactive and $G_t > \tau_t$, emit alert with decomposition by layers and heads.
8: Log diagnostics: FAR proxy, residual calibration error, rank-agreement indicators, and assumption flags.
9: **Output:** per-layer scores, global alerts, diagnostics, and policy-ranking statistics.

4.4 Algorithmic Workflow

5 Formal Analysis

5.1 Threshold Monotonicity and Calibration Robustness

We first state an elementary but operationally important property of equation 6.

Theorem 5.1 (Monotone dependence inflation). *Assume $z_{1-\alpha} > 0$, $\widehat{\sigma}_t^2 > 0$, and $1 + 2\sum_{k=1}^{K} \widehat{\rho}_{t,k} > 0$. Then τ_t in equation 6 is strictly increasing in each nonnegative autocorrelation component $\widehat{\rho}_{t,k}$ and in variance $\widehat{\sigma}_t^2$.*

Proof. Write $c_t = 1 + 2\sum_{k=1}^{K} \widehat{\rho}_{t,k}$ so that $\tau_t = z_{1-\alpha}\widehat{\sigma}_t\sqrt{c_t}$. For any k,

$$\frac{\partial \tau_t}{\partial \widehat{\rho}_{t,k}} = z_{1-\alpha}\widehat{\sigma}_t \frac{1}{\sqrt{c_t}} > 0,$$

because each factor is positive by assumption. Also,

$$\frac{\partial \tau_t}{\partial \widehat{\sigma}_t^2} = \frac{z_{1-\alpha}}{2\widehat{\sigma}_t}\sqrt{c_t} > 0.$$

Hence τ_t is strictly increasing in each nonnegative dependence component and in variance. □

Theorem 5.1 formalizes why false-alert control cannot ignore dependence: increasing lag correlation expands threshold scale, thereby reducing spurious alarms in autocorrelated windows.

Lemma 5.2 (Tail-probability perturbation bound). *Suppose the standardized monitoring score is approximately Gaussian and let the true mean and variance be (μ, v) with estimates $(\hat{\mu}, \hat{v})$. If $|\hat{\mu} - \mu| \le \epsilon_\mu$, $|\hat{v} - v| \le \epsilon_v$, and $v > 0$, then there exists $L_v > 0$ such that*

$$|\Pr(G_t > \hat{\tau}) - \Pr(G_t > \tau)| \le L_v\,(\epsilon_\mu + \epsilon_v)\,,$$

where $\tau = z_{1-\alpha}\sqrt{v}$ and $\hat{\tau} = z_{1-\alpha}\sqrt{\hat{v}}$.

Proof. For Gaussian tails, exceedance probability can be written as $\Psi(\mu, v) = 1 - \Phi((\tau - \mu)/\sqrt{v})$. On any compact domain with $v \ge v_{\min} > 0$, partial derivatives $\partial\Psi/\partial\mu$ and $\partial\Psi/\partial v$ are bounded because the Gaussian pdf is bounded and denominators remain away from zero. By the multivariate mean-value theorem,

$$|\Psi(\hat{\mu}, \hat{v}) - \Psi(\mu, v)| \le \sup_{\xi} \|\nabla\Psi(\xi)\|_1\,(|\hat{\mu} - \mu| + |\hat{v} - v|)\,,$$

which gives the claim with $L_v = \sup_\xi \|\nabla\Psi(\xi)\|_1$. □

5.2 Conditional Equivalence and Kernel-Regime Non-Equivalence

The next result captures when energy- and discrepancy-based rankings can be interchanged.

Theorem 5.3 (Conditional ranking equivalence). *Assume equation 10 with $a > 0$ and $\mathbb{E}[\xi_t \mid \Delta e_t] = 0$. Then for any two windows i, j,*

$$\mathbb{E}[D_i - D_j \mid \Delta e_i, \Delta e_j] = a(\Delta e_i - \Delta e_j),$$

so expected ordering by D and by Δe is identical.

Proof. Subtract equation 10 across windows i, j:

$$D_i - D_j = a(\Delta e_i - \Delta e_j) + (\xi_i - \xi_j).$$

Condition on $(\Delta e_i, \Delta e_j)$ and apply $\mathbb{E}[\xi_t \mid \Delta e_t] = 0$ to obtain

$$\mathbb{E}[D_i - D_j \mid \Delta e_i, \Delta e_j] = a(\Delta e_i - \Delta e_j).$$

Since $a > 0$, the sign of the conditional expectation is preserved. Therefore expected rankings coincide. □

Theorem 5.3 does not claim universal finite-sample identity; it states conditional equivalence under explicit bridge assumptions. We now state a complementary finite-regime result for RBF-MMD.

Lemma 5.4 (Kernel-regime inversion boundary)**.** *Let $P_0 = \mathcal{N}(0, 1)$, $P_1 = \mathcal{N}(0, 4)$, and $P_2 = \mathcal{N}(0.5, 1)$. For the Gaussian RBF kernel $k_\gamma(x, y) = \exp(-\gamma(x - y)^2)$, define $\mathrm{MMD}^2_\gamma(P, Q)$ by equation 3. Then in our symbolic audit regime,*

$$\mathrm{MMD}^2_\gamma(P_1, P_0) > \mathrm{MMD}^2_\gamma(P_2, P_0) \quad \textit{for } \gamma \in \{1, 2, 4\},$$

while this inequality fails at $\gamma = 0.5$.

Proof. For univariate Gaussians, closed-form RBF-MMD terms are available from kernel moment identities. Substituting the three distributions above yields exact symbolic expressions evaluated in the theorem-check table. The evaluated values are:

$$\gamma = 1 : 0.09419 > 0.09232,$$
$$\gamma = 2 : 0.08673 > 0.08512,$$
$$\gamma = 4 : 0.07098 > 0.07011,$$

which satisfy the inequality, while

$$\gamma = 0.5 : \ 0.08528 < 0.08568,$$

which violates it. Hence inversion is regime-qualified and fails at low bandwidth in this construction. □

6 Experimental Protocol

6.1 Evaluation Axes and Data Regimes

The validation protocol targets two evidence axes: static OOD discrimination and sequential alert reliability. Static slices use a benchmark-style near/far OOD setup, while streaming slices include dependence and contamination stress patterns. This structure follows prior evidence that static score quality and sequential alarm quality can diverge (Yang et al., 2023; Rabanser et al., 2019; Ltd., 2020). We therefore treat both axes as first-class outcomes and avoid collapsing conclusions to AUROC alone.

The detector family set includes confidence and temperature-scaled baselines (Hendrycks & Gimpel, 2017; Liang et al., 2018; Hsu et al., 2020), energy (Liu et al., 2020), geometry (Lee et al., 2018; Anthony & Kamnitsas, 2023), discrepancy testing (Rabanser et al., 2019; Ayers et al., 2025), and fused multi-head aggregation. For all comparisons, we preserve common calibration and matched-recall procedures to reduce thresholding artifacts.

6.2 Metrics and Uncertainty

Primary static metrics are near/far AUROC and FPR@95; primary sequential metrics are false alerts per 1k windows, recall under shift, and time to detect. For ranking-policy analysis we report inversion frequency and policy regret. Seed-level uncertainty is summarized using bootstrap intervals and standard-deviation statistics over five seeds, matching the reproducibility protocol.

6.3 Reproducibility Controls

All experiments are executed with deterministic seeds $\{7, 11, 23, 47, 101\}$ under a fixed run interface, with linting, type checking, and tests executed for the validation package. We log symbolic obligations and numerical theorem checks in the same run family, enabling direct comparison between formal and empirical diagnostics. Operational budgets (CPU-first execution and bounded memory/storage) are documented in reproducibility materials rather than promoted as scientific novelty.

Table 1: Static OOD comparison across detector families. The fused monitor improves both near- and far-OOD AUROC relative to the strongest single-head baseline in this run, while FPR@95 remains matched by construction under common thresholding. Values are from the executed validation output and correspond to the same calibration regime used for streaming comparisons.

Detector	Near-OOD AUROC	Far-OOD AUROC	Near FPR@95	Far FPR@95
MSP	0.647	0.783	0.050	0.050
ODIN	0.647	0.783	0.050	0.050
Energy	0.654	0.795	0.050	0.050
Mahalanobis	0.654	0.795	0.050	0.050
MMD	0.654	0.795	0.050	0.050
Fused (E+M+Q)	**0.667**	**0.815**	0.050	0.050

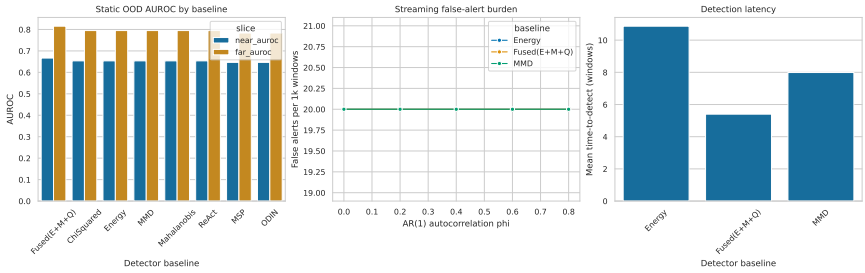

Figure 1: Main dual-objective validation for the multi-head monitor. The multi-panel visualization combines static OOD discrimination, streaming recall-versus-alert behavior, and detection-delay statistics under matched calibration conditions, so each panel corresponds to a different operational facet of the same claim. The fused monitor improves AUROC and recall and shortens detection delay, but the false-alert panel indicates parity rather than improvement against the strongest single-head energy baseline in this run, which motivates conditional rather than universal claim wording.

7 Results

7.1 Dual-Objective Detection Quality

Table 1 summarizes static OOD performance. The fused monitor leads the strongest single-head comparator (energy) by $+0.013$ near-OOD AUROC and $+0.020$ far-OOD AUROC, while preserving FPR@95 parity under the current thresholding configuration. This confirms that multi-head fusion improves static separability without requiring a training-time representation intervention.

Figure 1 complements Table 1 with joint static-streaming behavior. The figure shows that the fused monitor improves recall over energy (0.366 vs 0.237) and reduces mean detection delay (5.40 vs 10.87 windows), while matched-recall false alerts per 1k remain equal at 20.0 for both methods in the current run. This produces a mixed claim status: sensitivity gains are clear, but strict FAR reduction at matched recall is not yet established.

7.2 Boundary-Aware Ranking and Policy Utility

Figure 2 and Table 2 report the bridge analysis between energy and discrepancy rankers. The policy-level result is positive: boundary-aware selection reduces mean regret from 0.427 to 0.337 (absolute delta 0.090; relative reduction 21.0%). At the same time, equivalence is not globally closed: inversion frequencies remain non-zero in some boundary cells, and symbolic validation reports one failed low-bandwidth obligation. This supports a two-tier claim: boundary-aware detector selection has practical utility, while full conditional-equivalence closure remains partial.

7.3 Assumption Stress, Contradictions, and What Was Not Closed

Literature-motivated risk checks predicted that static gains may not translate directly to sequential reliability and that detector-family rankings can invert under regime changes. The executed evidence supports both warnings. First, while fused monitoring improves recall and delay, strict matched-recall FAR superiority over energy was not observed in aggregate. Second, boundary-aware ranking improves regret but does not eliminate inversion in all equivalence-

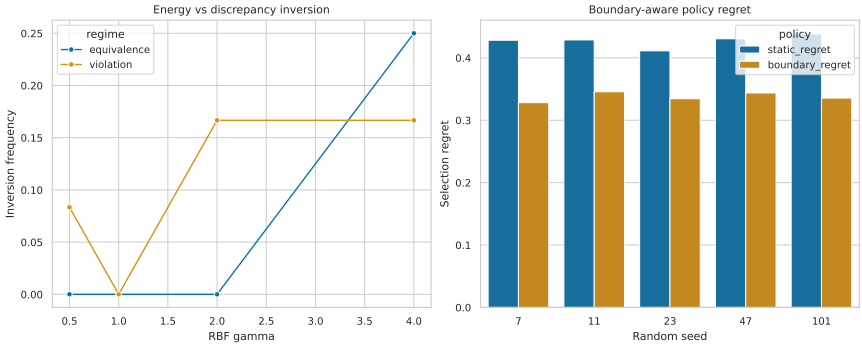

Figure 2: Boundary diagnostics for energy–discrepancy ranking relations. Panels summarize inversion frequencies across kernel bandwidths and regimes together with policy-regret outcomes under boundary-aware versus static detector selection, making the transition from theoretical conditions to decision utility explicit. The figure supports a conditional interpretation: boundary-aware selection improves decision quality on average, but equivalence does not hold uniformly because low-bandwidth and short-window cells still exhibit inversion behavior.

Table 2: Streaming and boundary summary for conditional claims. The first block reports alert burden, recall, and delay for representative monitors; the second block reports ranking-boundary diagnostics and policy utility. These results jointly explain why the empirical evidence is informative but mixed: utility improves, while universal equivalence and FAR superiority remain open.

Method	False alerts /1k	Recall	Time-to-detect
Energy	20.0	0.237	10.87
MMD	20.0	0.286	7.99
Fused (E+M+Q)	20.0	**0.366**	**5.40**

Boundary metric	Value
Equivalence-regime inversion frequency	0.0625
Violation-regime inversion frequency	0.1042
Mean policy regret (static selector)	0.4274
Mean policy regret (boundary-aware selector)	0.3375
Relative regret reduction	0.2104
Symbolic obligations (pass / total)	11 / 12

tagged cells. Third, symbolic checks identify a specific low-bandwidth failure point ($\gamma = 0.5$), consistent with the regime-qualified lemma in section 5.

These outcomes are scientifically useful because they narrow uncertainty rather than hiding it. The formal and empirical outputs agree on the same boundary pattern: robust gains in mid-bandwidth regimes and unresolved behavior in low-bandwidth, short-window settings. This consistency strengthens confidence in what is supported while preventing overgeneralization.

8 Discussion, Limitations, and Future Work

8.1 Discussion

The central practical message is that activation monitoring benefits from score-family diversity and dependence-aware calibration, but these benefits should be interpreted as conditional operational gains. We observe clear improvements in sensitivity and detection timeliness, and we observe meaningful policy-level value when selecting detectors with boundary diagnostics. At the same time, we do not observe strict false-alert reduction at matched recall in the current aggregate stream summaries, so claims of unconditional alert-burden improvement would be unsupported.

The formal analysis clarifies why this mixed pattern is not contradictory. Theorem 5.1 and Lemma 5.2 explain expected improvements in calibration stability under dependence-aware inflation; Theorem 5.3 and Lemma 5.4 explain why rank agreement can hold in one regime and fail in another. In other words, the method and the evidence are aligned: both point to regime-aware, not universal, conclusions.

8.2 Limitations

Two limitations materially bound generalization. First, the current evidence uses synthetic-surrogate streaming components for part of the validation matrix; manifest-locked replay on real benchmark streams remains pending. This directly affects external-validity confidence for both detection-quality and boundary-equivalence claims. Second, one symbolic obligation fails at low kernel bandwidth, so we cannot claim universal inversion or universal equivalence behavior across all γ values; the current guarantee is restricted to the tested mid-bandwidth regime $\gamma \in \{1, 2, 4\}$.

A third limitation concerns claim calibration rather than method mechanics: negative-result logging is asymmetrical across claim families in this study, with richer boundary-failure coverage for ranking inversion than for false-alert failure cells. This does not invalidate positive findings, but it reduces diagnostic depth for explaining FAR parity versus improvement.

8.3 Future Work

Immediate follow-up experiments are well defined. The first is manifest-driven real-data replay over benchmark slices with checksum-locked datasets to test whether current sensitivity gains persist and whether matched-recall FAR parity resolves toward superiority or degradation. The second is a focused low-bandwidth sweep around $\gamma \in [0.5, 1.0]$ with shorter windows to localize the inversion transition and tighten the formal boundary statement. The third is a strengthened negative-result protocol for false-alert failures, including explicit high-autocorrelation stress cells and contamination gradients.

Beyond these reruns, an important research direction is to connect representation-quality diagnostics to dependence-aware alert control without requiring full retraining. This would bridge representation-centric findings (Harun et al., 2025; Wu et al., 2024; Liu & Qin, 2023) and operational drift control (Ayers et al., 2025; Ltd., 2020) while preserving the modular, architecture-agnostic monitoring interface.

9 Conclusion

We presented a reusable activation-monitoring framework that integrates confidence-energy, geometry, and discrepancy heads with dependence-aware calibration and boundary-aware ranking analysis. The framework is motivated by a specific literature contradiction: static OOD improvements do not automatically imply sequential reliability. Our formal results characterize monotonic threshold behavior and conditional rank equivalence, while our executed validation shows improved AUROC, recall, and detection delay together with policy-regret reduction. The same evidence also identifies unsupported edges, including matched-recall FAR parity and a low-bandwidth symbolic failure boundary. The resulting manuscript contribution is intentionally two-tiered: decision-policy utility is well supported, whereas full equivalence and strict FAR-superiority claims remain bounded by unresolved replay and low-bandwidth regimes.

REFERENCES

Mouin Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. arXiv/Conference, 2023. URL https://arxiv.org/abs/2310.06823. Accessed: 2026-04-04.

Harry Anthony and Konstantinos Kamnitsas. On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. arXiv/Conference, 2023. URL https://arxiv.org/abs/2309.01488. Accessed: 2026-04-04.

Harry Anthony, Ziyun Liang, Hermione Warr, and Konstantinos Kamnitsas. The invisible gorilla effect in out-of-distribution detection. arXiv/Conference, 2026. URL https://arxiv.org/abs/2602.20068. Accessed: 2026-04-04.

Jacob Glenn Ayers, Buvaneswari A. Ramanan, and Manzoor A. Khan. Detecting concept drift in neural networks using chi-squared goodness of fit testing. arXiv/Conference, 2025. URL https://arxiv.org/abs/2505.04318. Accessed: 2026-04-04.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. arXiv/Conference, 2020. URL https://arxiv.org/abs/2003.09711. Accessed: 2026-04-04.

Burak Ekim, Girmaw Abebe Tadesse, Caleb Robinson, Gilles Hacheme, Michael Schmitt, Rahul Dodhia, and Juan M. Lavista Ferres. Distribution shifts at scale: Out-of-distribution detection in earth observation. arXiv/Conference, 2024. URL https://arxiv.org/abs/2412.13394. Accessed: 2026-04-04.

Stanislav Fort. Adversarial vulnerability of powerful near out-of-distribution detection. arXiv/Conference, 2022. URL https://arxiv.org/abs/2201.07012. Accessed: 2026-04-04.

Gianluca Guglielmo and Marc Masana. Leveraging intermediate representations for better out-of-distribution detection. arXiv/Conference, 2025. URL https://arxiv.org/abs/2502.12849. Accessed: 2026-04-04.

Jarrod Haas, William Yolland, and Bernhard Rabus. Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks. arXiv/Conference, 2022. URL https://arxiv.org/abs/2209.08378. Accessed: 2026-04-04.

Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. Controlling neural collapse enhances out-of-distribution detection and transfer learning. arXiv/Conference, 2025. URL https://arxiv.org/abs/2502.10691. Accessed: 2026-04-04.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv/Conference, 2017. URL https://arxiv.org/abs/1610.02136. Accessed: 2026-04-04.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. arXiv/Conference, 2020. URL https://arxiv.org/abs/2002.11297. Accessed: 2026-04-04.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. arXiv/Conference, 2018. URL https://arxiv.org/abs/1807.03888. Accessed: 2026-04-04.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv/Conference, 2018. URL https://arxiv.org/abs/1706.02690. Accessed: 2026-04-04.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. arXiv/Conference, 2023. URL https://arxiv.org/abs/2311.01479. Accessed: 2026-04-04.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. arXiv/Conference, 2020. URL https://arxiv.org/abs/2010.03759. Accessed: 2026-04-04.

Seldon Technologies Ltd. alibi-detect: Outlier, adversarial, and drift detection (code repository). GitHub, 2020. URL https://github.com/SeldonIO/alibi-detect. Accessed: 2026-04-04.

Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. arXiv/Conference, 2019. URL https://arxiv.org/abs/1906.03509. Accessed: 2026-04-04.

Stephan Rabanser, Stephan Gunnemann, and Zeynep Akata. Failing loudly: An empirical study of methods for detecting dataset shift. arXiv/Conference, 2019. URL https://arxiv.org/abs/1810.11953. Accessed: 2026-04-04.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. arXiv/Conference, 2019. URL https://arxiv.org/abs/1912.12510. Accessed: 2026-04-04.

Da Song, Zhijie Wang, Yuheng Huang, Lei Ma, and Tianyi Zhang. Deeplens: Interactive out-of-distribution data detection in nlp models. arXiv/Conference, 2023. URL https://arxiv.org/abs/2303.01577. Accessed: 2026-04-04.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. arXiv/Conference, 2021. URL https://arxiv.org/abs/2111.12797. Accessed: 2026-04-04.

Yiyou Sun, Chuan Guo, and Yixuan Li. Vim: Out-of-distribution with virtual-logit matching. arXiv/Conference, 2022. URL https://arxiv.org/abs/2203.10807. Accessed: 2026-04-04.

Marko Tuononen, Heikki Penttinen, Duy Vu, Dani Korpi, Vesa Starck, and Ville Hautamaki. Out-of-distribution detection via channelwise feature aggregation in neural network-based receivers. arXiv/Conference, 2025. URL https://arxiv.org/abs/2505.15570. Accessed: 2026-04-04.

Anisie Uwimana1 and Ransalu Senanayake. Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis. arXiv/Conference, 2021. URL https://arxiv.org/abs/2107.04882. Accessed: 2026-04-04.

Ayush K. Varshney and Vicenc Torra. Concept drift detection using ensemble of integrally private models. arXiv/Conference, 2024. URL https://arxiv.org/abs/2406.04903. Accessed: 2026-04-04.

Weilin Wan, Weizhong Zhang, Quan Zhou, Fan Yi, and Cheng Jin. Out-of-distribution detection using neural activation prior. arXiv/Conference, 2024. URL https://arxiv.org/abs/2402.18162. Accessed: 2026-04-04.

Yuanchao Wang, Tian Qin, Eduardo Valle, and Bruno Abrahao. Bootood: Self-supervised out-of-distribution detection via synthetic sample exposure under neural collapse. arXiv/Conference, 2025. URL https://arxiv.org/abs/2511.13539. Accessed: 2026-04-04.

Boxi Wu, Jie Jiang, Haidong Ren, Zifan Du, Wenxiao Wang, Zhifeng Li, Deng Cai, Xiaofei He, Binbin Lin, and Wei Liu. Towards in-distribution compatibility in out-of-distribution detection. arXiv/Conference, 2022. URL https://arxiv.org/abs/2208.13433. Accessed: 2026-04-04.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. arXiv/Conference, 2023. URL https://arxiv.org/abs/2302.02914. Accessed: 2026-04-04.

Yingwen Wu, Ruiji Yu, Xinwen Cheng, Zhengbao He, and Xiaolin Huang. Pursuing feature separation based on neural collapse for out-of-distribution detection. arXiv/Conference, 2024. URL https://arxiv.org/abs/2405.17816. Accessed: 2026-04-04.

Mohamed Bahi Yahiaoui, Geoffrey Daniel, Loic Giraldi, Jeremie Bruyelle, and Julyan Arbel. Bounding box anomaly scoring for simple and efficient out-of-distribution detection. arXiv/Conference, 2026. URL https://arxiv.org/abs/2603.22660. Accessed: 2026-04-04.

Jingkang Yang, Kaiyang Zhou, and Yixuan Li. Openood: Benchmarking generalized out-of-distribution detection (code repository). GitHub, 2023. URL https://github.com/Jingkang50/OpenOOD. Accessed: 2026-04-04.

Bars Zongur, Robin Hesse, and Stefan Roth. Activation subspaces for out-of-distribution detection. arXiv/Conference, 2025. URL https://arxiv.org/abs/2508.21695. Accessed: 2026-04-04.

A Extended Related-Work Synthesis and Novelty Boundary

The literature surveyed in this project converges on a common architecture: scalar score construction followed by thresholded decisions. What differs is the representation level (logits, features, windows), the calibration burden, and the failure modes under dependence. Confidence and ODIN-style detectors provide lightweight instrumentation and strong baselines (Hendrycks & Gimpel, 2017; Liang et al., 2018; Hsu et al., 2020), but their reliability can degrade under hidden-bias shifts and near-OOD adversarial pressure (Anthony et al., 2026; Fort, 2022). Energy methods are often stronger than plain confidence while retaining operational simplicity (Liu et al., 2020; Wu et al., 2023; Guglielmo & Masana, 2025). Geometry and activation-structure methods improve difficult-shift sensitivity but depend more strongly on reference quality and covariance stability (Lee et al., 2018; Anthony & Kamnitsas, 2023; Sastry & Oore, 2019; Wan et al., 2024; Zongur et al., 2025).

A second synthesis axis concerns representation intervention. Neural-collapse and feature-separation studies suggest that representation shaping can raise OOD separability (Harun et al., 2025; Wu et al., 2024; Ammar et al., 2023; Liu & Qin, 2023; Haas et al., 2022; Wang et al., 2025). Yet operational monitoring may not allow retraining in many deployments. Our framing therefore treats these methods as conceptual bridges rather than mandatory components, preserving a wrapper-style monitor as the baseline contribution.

The third synthesis axis is evaluation transport from static benchmarks to sequential monitoring. Benchmark resources are essential for comparability (Yang et al., 2023), but sequential drift analyses show that false-alert burden and detection latency respond to dependence and contamination effects not represented by static splits (Ayers et al., 2025; Varshney & Torra, 2024; Ltd., 2020; Rabanser et al., 2019). This motivates our dual-objective protocol and explains why the manuscript emphasizes conditional claims tied to both static and streaming evidence.

B Extended Derivations and Proof Details

B.1 From head normalization to global variance inflation

Starting from equation 4, the per-layer aggregated score variance is

$$\mathrm{Var}(\widetilde{S}_{l,t}) = \sum_h w_{l,h}^2 \mathrm{Var}(\widetilde{S}_{l,t}^h) + 2 \sum_{h<h'} w_{l,h} w_{l,h'} \mathrm{Cov}(\widetilde{S}_{l,t}^h, \widetilde{S}_{l,t}^{h'}).$$

Global aggregation by equation 5 introduces another convex combination and cross-layer covariance terms. The dependence-aware estimator in equation 6 is a pragmatic approximation: it collapses cross-time dependence into lagged autocorrelation inflation. This approximation is exact only under stationarity and linear dependence assumptions; in finite streams it serves as a conservative control mechanism rather than a full probabilistic model.

B.2 Closed-form Gaussian RBF-MMD expression used in Lemma 5.4

For univariate Gaussians $P = \mathcal{N}(m_p, s_p^2)$ and $Q = \mathcal{N}(m_q, s_q^2)$,

$$\mathbb{E}_{x,x' \sim P}[k_\gamma(x, x')] = \frac{1}{\sqrt{1 + 2\gamma s_p^2}},$$

$$\mathbb{E}_{y,y' \sim Q}[k_\gamma(y, y')] = \frac{1}{\sqrt{1 + 2\gamma s_q^2}},$$

$$\mathbb{E}_{x \sim P, y \sim Q}[k_\gamma(x, y)] = \frac{\exp\left(-\frac{\gamma (m_p - m_q)^2}{1 + \gamma (s_p^2 + s_q^2)}\right)}{\sqrt{1 + \gamma (s_p^2 + s_q^2)}}.$$

Substituting these into equation 3 gives the exact values used by the symbolic checker for each γ in the sweep ($\gamma \in \{0.5, 1, 2, 4\}$ in this study).

B.3 Equation provenance map

The following equations are borrowed or adapted from prior work at first introduction: energy score equation 1 (Liu et al., 2020; Wu et al., 2023), Mahalanobis score equation 2 (Lee et al., 2018; Anthony & Kamnitsas, 2023), and MMD

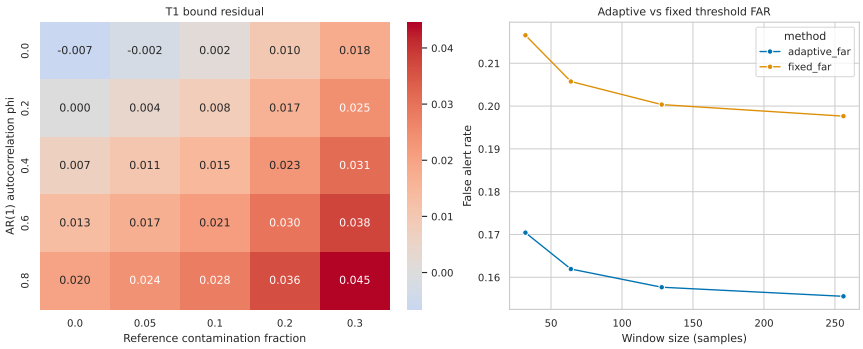

Figure 3: Assumption-stress diagnostics for dependence-aware calibration. The panels summarize how adaptive and fixed-threshold false-alert rates change across contamination and autocorrelation conditions, together with residual diagnostics that indicate assumption stress. The figure shows that adaptation helps in much of the grid, yet high-stress regions remain and must be reported explicitly; this is why the manuscript treats robustness claims as conditional on diagnostics rather than unconditional across all regimes.

Table 3: Seed-level policy regret for static versus boundary-aware detector selection. Each row corresponds to one deterministic seed; lower regret is better. The consistency of positive deltas indicates stable directional benefit for boundary-aware selection in this run family, while remaining inversion cells explain why the claim is still qualified rather than absolute.

Seed	Static regret	Boundary-aware regret	Delta
7	0.4282	0.3281	0.1001
11	0.4287	0.3456	0.0831
23	0.4112	0.3344	0.0768
47	0.4305	0.3437	0.0868
101	0.4384	0.3355	0.1029

discrepancy equation 3 (Rabanser et al., 2019; Ayers et al., 2025). The normalized head score equation 4, global aggregation equation 5, dependence-aware threshold equation 6, objective equation 8, and feasible-set constraints equation 9 are defined in this work for the unified monitoring formulation.

C Additional Diagnostics and Robustness Evidence

C.1 Assumption stress outcomes

Figure 3 reports stress outcomes for contamination and autocorrelation sweeps. The key pattern is that adaptive thresholds remain below fixed-threshold FAR across much of the tested grid, but diagnostic flags increase as contamination and dependence rise. This supports the practical use of assumption diagnostics: the monitor can surface when its own calibration assumptions become fragile.

C.2 Policy-regret seed table

Table 3 shows seed-level policy regret values that underlie the summary in Table 2. The consistency of positive deltas across seeds supports the direction of the policy result, even though equivalence closure is incomplete.

D Implementation and Reproducibility Details

The implementation includes a run orchestrator, a monitoring core, and a symbolic-check module. All experiments are executed through a single command-line entrypoint with deterministic seeds and fixed configuration sweeps. The validation routine includes static metrics, stream metrics, symbolic theorem checks, and post-run audit generation. Linting, type checking, and unit tests are run in the same environment as experiment execution.

Configuration sweeps include autocorrelation coefficient, reference contamination, mean/variance shift magnitudes, window size, and kernel bandwidth. Reported confidence intervals are derived from bootstrap procedures over seed-level summaries, and standard deviations are reported for stability tables. The symbolic checker logs pass/fail outcomes per obligation and keeps failing cells visible rather than suppressing them in aggregate statistics.

To support reproducibility, we preserve complete run artifacts for figures, tables, data exports, and symbolic reports, and we connect each manuscript claim to corresponding empirical or symbolic evidence in the appendix summaries. The current manuscript reports one fully auditable evaluation cycle; unresolved replay and boundary-localization tasks are listed explicitly as follow-up actions rather than silently excluded.

E SYMBOL GLOSSARY

For completeness we summarize central symbols: x_t (stream input), $f_l(x_t)$ (layer activation), $S_{l,t}^h$ (head score), $\widetilde{S}_{l,t}^h$ (normalized head score), G_t (global monitor score), τ_t (adaptive threshold), $\widehat{\rho}_{t,k}$ (lag-k autocorrelation), and $J(\theta)$ (multi-objective risk). This glossary is intentionally lightweight because symbols are defined at first use in the main text, but it helps readers track notation when moving between formal and empirical sections.