STREAM-OOD: Regime-Aware Sequential Monitoring for Streaming Out-of-Distribution Detection

Boshra Khalili
Department of Civil Engineering and Engineering Mechanics
Columbia University
New York, NY, USA
bk2898@columbia.edu

Alec Bardey
New York City Department of Transportation
New York, NY, USA
abardey@dot.nyc.gov

Andrew W. Smyth
Department of Civil Engineering and Engineering Mechanics
Columbia University
New York, NY, USA
aws16@columbia.edu

Appendix

A. Sequential Detection Interpretation with Adaptive Hazard

A.1. Streaming Mean-Shift Formulation

After standardization,

$$Z_t = \frac{S_t - \mu_{ID}}{\sigma_{ID}}, \tag{1}$$

we model the novelty score stream $\{Z_t\}$ as a piecewise stochastic process with an unknown change time $\tau^\star$:

$$Z_t \sim \begin{cases} \mathcal{D}_0, & t < \tau^\star, \\ \mathcal{D}_1, & t \geq \tau^\star. \end{cases} \tag{2}$$

We assume:

- $\mathbb{E}_{\mathcal{D}_0}[Z_t] = 0$ (ID regime),
- $\mathbb{E}_{\mathcal{D}_1}[Z_t] = \delta > 0$ (sustained OOD regime),
- $\mathcal{D}_0$ and $\mathcal{D}_1$ are sub-Gaussian with variance proxy σ^2.

Under this formulation, sustained OOD corresponds to a persistent positive mean shift in the standardized novelty stream. Streaming OOD monitoring therefore reduces to a quickest change detection problem under a potentially time-varying regime transition prior.

This connection situates STREAM-OOD within the classical change detection framework while relaxing the constant-hazard assumption to allow context-dependent transition dynamics.

A.2. Adaptive Hazard Regime Model

Let r_t denote the run-length (time since the most recent regime change). In classical Bayesian Online Change Point Detection (BOCPD), regime transitions occur with a constant hazard rate H.

In STREAM-OOD, we instead introduce an adaptive hazard:

$$H_t = P(r_t = 0 \mid \phi_t), \tag{3}$$

where ϕ_t is a streaming feature vector constructed from novelty statistics (e.g., Z_t, ΔZ_t, $\mathrm{EMA}(Z_t)$, and gating state).

The run-length recursion becomes:

$$P(r_t \mid Z_{1:t}) = \begin{cases} \displaystyle\sum_{r_{t-1}} P(r_{t-1} \mid Z_{1:t-1})\, H_t\, P(Z_t \mid r_t = 0), \\ r_t = 0, \\ P(r_{t-1} = r_t - 1 \mid Z_{1:t-1})(1 - H_t) P(Z_t \mid r_t), \\ r_t > 0. \end{cases} \tag{4}$$

When $H_t \equiv H$ is constant, Eq. (4) reduces to classical BOCPD. Adaptive hazard modeling therefore generalizes constant-hazard sequential detection by allowing transition probability to depend on streaming evidence.

A.3. Information-Theoretic Detection Limit

Define the Kullback–Leibler divergence between regimes:

$$D(\mathcal{D}_1 \| \mathcal{D}_0) = \mathbb{E}_{\mathcal{D}_1}\left[\log \frac{f_1(Z)}{f_0(Z)}\right]. \tag{5}$$

Under mild regularity conditions, any stopping rule controlling false-alarm probability ϵ satisfies

$$\mathbb{E}[\mathrm{TTD}] \geq \frac{\log(1/\epsilon)}{D(\mathcal{D}_1 \| \mathcal{D}_0)}, \quad \epsilon \to 0. \tag{6}$$

This lower bound depends solely on the divergence between regimes and is independent of the specific hazard parameterization. Consequently, adaptive hazard modeling does not alter the fundamental delay–false alarm limit; rather, it modulates how evidence is accumulated relative to contextual streaming statistics while remaining subject to the same information-theoretic constraints.

A.4. Event-Level Metric Definitions

Let the set of ground-truth OOD events be $\mathcal{E} = \{E_k\}_{k=1}^K$, where each event is a contiguous interval

$$E_k = [s_k, e_k],$$

and let the set of predicted alert segments be $\hat{\mathcal{E}} = \{\hat{E}_j\}_{j=1}^J$, where

$$\hat{E}_j = [\hat{s}_j, \hat{e}_j].$$

Let $|\cdot|$ denote interval duration in frames, and define the overlap indicator

$$O_{kj} = \mathbf{1}[E_k \cap \hat{E}_j \neq \emptyset].$$

Time-to-Detect (TTD). For each ground-truth event E_k, we define the detection delay as the number of frames between the event onset s_k and the first predicted alert that overlaps the event:

$$d_k = \min\Big(\{t - s_k :\ t \in E_k \cap (\cup_j \hat{E}_j)\} \cup \{|E_k|\}\Big).$$

If an event is never detected, we assign the maximum delay $|E_k|$. The overall Time-to-Detect is

$$\mathrm{TTD} = \frac{1}{K}\sum_{k=1}^{K} d_k.$$

False Alarms per Hour (FA/hr). A predicted alert segment is counted as a false alarm if it does not overlap any ground-truth OOD event. Let T_{ID} denote the total duration of in-distribution-only video in seconds. Then

$$\text{FA/hr} = \frac{\sum_{j=1}^{J} \mathbf{1}\left[\sum_{k=1}^{K} O_{kj} = 0\right]}{T_{\text{ID}}/3600}.$$

Fragmentation (Frag). Fragmentation measures how many distinct predicted alert segments overlap a single true event, averaged across all events:

$$\text{Frag} = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} O_{kj}.$$

Lower values indicate more coherent event-level alerts, with the ideal case approaching one alert segment per true event.

Stability. Stability measures the fraction of each ground-truth event that is covered by predicted alerts:

$$\text{Stab} = \frac{1}{K} \sum_{k=1}^{K} \frac{|E_k \cap (\cup_j \hat{E}_j)|}{|E_k|}.$$

Higher values indicate more continuous alert coverage throughout the event duration.

These metrics complement AUROC and FPR@95 by explicitly quantifying detection latency, false-alarm burden, alert fragmentation, and temporal coverage under realistic streaming correlation.

B. Training of Adaptive Hazard

While the adaptive hazard provides a probabilistic formulation of regime transitions, its practical effectiveness depends on how transition probabilities are estimated from streaming statistics. We therefore describe the supervised protocol used to learn the hazard function under strict causal deployment constraints.

Parameterization. The adaptive hazard is defined as

$$H_t = \sigma\big(g_\theta(\phi_t)\big), \tag{7}$$

where $g_\theta(\cdot)$ is a lightweight two-layer multilayer perceptron (hidden dimension 32), θ denotes learnable parameters, and $\sigma(\cdot)$ is the sigmoid function. The hazard outputs the probability of a regime transition at time t, conditioned only on past and present statistics.

Supervision Protocol. We train the hazard network using event-level regime supervision. For each time step t, we define a binary transition label $y_t \in \{0, 1\}$ indicating whether a regime boundary occurs at time t. Importantly, supervision is applied exclusively to regime transitions rather than to frame-level OOD classification. This ensures that the hazard models temporal transition dynamics instead of learning semantic discriminators tied to specific visual content.

The hazard parameters are optimized using binary cross-entropy:

$$\mathcal{L}_{\text{hazard}} = -y_t \log H_t - (1 - y_t) \log(1 - H_t). \tag{8}$$

Causality and Information Isolation. Training is performed strictly causally. The streaming feature vector

$$\phi_t = \big[Z_t,\ \Delta Z_t,\ \text{EMA}(Z_t),\ g_t\big] \tag{9}$$

is computed using only information available up to time t, eliminating hindsight bias. Here, $\Delta Z_t = Z_t - Z_{t-1}$ captures score velocity, $\text{EMA}(Z_t)$ encodes low-frequency drift, and g_t denotes the conformal gating state.

To prevent representation leakage, the embedding extractor and the nonparametric manifold estimator remain frozen during hazard training. Optimization is therefore confined strictly to modeling transition dynamics in the standardized novelty stream, preserving representation independence and preventing feature-level overfitting.

Regularization and Capacity Control. The hazard network is intentionally lightweight to limit model capacity and discourage memorization of intersection-specific patterns. We apply ℓ_2 weight decay and early stopping based on validation streams. Because the hazard operates over standardized novelty statistics rather than raw visual features, its input dimensionality is low, further constraining expressiveness and promoting generalization.

Effect of Hazard Modeling and Sequential Baselines. Table 1 isolates the contribution of hazard modeling and compares it to classical sequential change detection. We include CUSUM as a modern quickest change detection baseline operating on the standardized novelty stream.

CUSUM reduces detection delay relative to frame-wise thresholding but exhibits elevated false alarms and fragmentation under non-stationary streaming conditions. Because CUSUM assumes a fixed post-change distribution and constant accumulation rate, it cannot adapt transition sensitivity to contextual streaming statistics such as score velocity or gating state.

A fixed constant hazard within the Bayesian framework improves stability relative to CUSUM but still suffers from

Method	TTD↓	FA/hr↓	Frag↓
CUSUM	3.5	2.3	2.0
Fixed Hazard ($H = 0.01$)	3.8	2.2	1.9
Fixed Tuned Hazard	3.3	2.1	1.7
Learned Hazard (MLP)	**3.0**	**1.6**	**1.3**

Table 1. Effect of sequential modeling strategies on event-level monitoring. Classical CUSUM improves delay but increases false alarms and fragmentation under streaming non-stationarity. Learning a context-dependent hazard weakens the delay–false-alarm coupling and yields the strongest monitoring reliability.

an intrinsic delay–false-alarm coupling. Tuning the constant hazard reduces delay modestly, yet fragmentation and false alarms remain elevated due to the inability to modulate transition probability dynamically.

Learning the hazard yields the strongest performance across all metrics. Detection delay decreases from 3.8 (fixed hazard) and 3.5 (CUSUM) to 3.0, false alarms reduce from 2.3 and 2.2 per hour to 1.6 (27% reduction), and fragmentation decreases from 2.0–1.9 to 1.3. These results demonstrate that context-dependent transition modeling weakens the intrinsic delay–false-alarm coupling observed under constant-hazard and classical sequential inference.

Cross-Location Generalization. To assess whether the learned hazard overfits to specific scene layouts or intersection geometry, we conduct a strict cross-location split. The hazard network is trained on streams from 20 intersections and evaluated on 10 disjoint intersections with no shared viewpoints or spatial overlap.

Table 2 reports cross-location performance. Detection delay increases modestly from 3.0 to 3.2 (+6.7%), while false alarms rise slightly from 1.6 to 1.7 per hour (+6.3%). Fragmentation increases marginally, and stability remains high (0.89 to 0.87).

The limited degradation across all metrics indicates that the hazard does not memorize location-specific appearance cues. Instead, it learns regime-transition dynamics over standardized novelty statistics that generalize across heterogeneous urban scenes. Because embeddings and manifold estimation remain frozen, performance on unseen intersections reflects robustness of sequential regime modeling rather than adaptation of visual representations.

Overall, adaptive hazard learning models regime-transition structure in the streaming novelty process while preserving causal deployment constraints, representation independence, and spatial generalization across heterogeneous urban camera networks.

Setting	TTD↓	FA/hr↓	Frag↓	Stab↑
Seen Intersections (20)	3.0	1.6	1.3	0.89
Unseen Intersections (10)	3.2	1.7	1.4	0.87
Relative Change (%)	+6.7%	+6.3%	+7.7%	-2.2%

Table 2. Cross-location generalization of the learned adaptive hazard. The hazard is trained on 20 intersections and evaluated on 10 unseen locations. Performance degradation remains minimal, indicating that the learned transition dynamics capture regime statistics rather than location-specific semantics.

C. Dataset Characteristics and Shift Statistics

Our dataset comprises 1,440 hours of continuous traffic-camera video collected from 30 heterogeneous New York City intersections under real deployment conditions. The streams cover diverse viewpoints and environmental conditions, including rain, snow, nighttime operation, partial visibility, and camera motion. To systematically evaluate robustness to structural and localized shifts, we further augment the real streams with controlled perturbations including partial lens occlusion and camera tilt. The dataset is private and not currently publicly releasable; we therefore provide detailed dataset statistics and evaluation protocols to support reproducibility.

Of the full corpus, 701 hours correspond to sustained regime shifts and 739 hours reflect normal in-distribution (ID) operation. Across 679 annotated OOD events, durations range from short localized perturbations lasting approximately 3 minutes (e.g., partial occlusion) to prolonged structural changes exceeding 12 hours (e.g., construction setup), with inter-event ID intervals frequently exceeding 5 hours. This temporal structure supports evaluation under realistic long-horizon monitoring constraints rather than short, artificially segmented clips.

Unlike conventional OOD benchmarks that assume independently sampled test images, this dataset reflects temporally correlated deployment conditions encountered in real-world monitoring systems. We curate deployment-realistic shifts spanning multiple forms of distribution change: global appearance changes due to sustained weather conditions, structural perturbations from viewpoint misalignment, localized corruption affecting object visibility, semantic novelty arising from rare vehicles (e.g., forklifts or construction machinery), and persistent scene evolution introduced by construction setup such as scaffolding or temporary barriers. Collectively, these scenarios span global, structural, semantic, and long-horizon scene-level deviations relevant to urban camera networks used for traffic counting, congestion estimation, and safety monitoring.

Structural shifts such as construction setup exhibit the longest persistence, often spanning multiple hours and cor-

responding to sustained geometric reconfiguration, occlusion changes, and altered traffic-flow patterns. Appearance-driven shifts such as sustained rain extend across illumination cycles, modifying global reflectance and texture statistics while largely preserving semantic scene layout. Camera tilt and partial lens occlusion introduce geometric and localized perturbations of intermediate duration.

The in-distribution traffic manifold is semantically rich: normal operation includes six vehicle categories—commercial vehicles, taxis, buses, school buses, trucks, and standard passenger vehicles—capturing realistic modal diversity in urban deployment. This prevents trivial novelty detection based on uncommon vehicle types and ensures that regime detection must rely on structured distributional shifts rather than simple semantic rarity.

All streams are processed strictly sequentially without access to future frames, enforcing real-time constraints and eliminating hindsight bias. Ground-truth OOD annotations are defined at the event level, where each event corresponds to a contiguous regime-shift interval. Because regime shifts are temporally continuous rather than frame-level anomalies, event durations span multiple temporal scales, from minutes to half-day structural changes, producing sustained mean shifts in representation space rather than isolated outliers. This enables rigorous evaluation of detection latency (TTD), alert fragmentation, false-alarm regulation, stability, and contamination risk under streaming correlation.

Collectively, these characteristics establish the dataset as a deployment-aligned benchmark for streaming OOD monitoring, where regime persistence, manifold stability, and event-level reliability are primary evaluation criteria rather than frame-level separability.

C.1. Baseline Details

We compare against representative representation-based OOD postprocessors operating on fixed CLIP embeddings, including MSP [2], ODIN [3], Energy-Based Scoring (EBO) [4], kNN Distance (KLM) [8], RMDS (RMD) [5], ViM [9], ReAct [7], ASH [1], and DICE [6]. These methods compute representation-level novelty scores without modifying the encoder or requiring retraining. We restrict comparison to frozen-embedding postprocessors because the goal of this work is to isolate the contribution of streaming inference, contamination-controlled adaptation, and regime modeling, rather than improvements due to representation learning or test-time encoder adaptation.

For streaming evaluation, we consider progressively stronger inference mechanisms. Static thresholding performs independent frame-wise decisions. EMA smoothing reduces high-frequency score oscillations. Bayesian online change-point inference operates on the standardized novelty score stream to model persistence under temporal correlation. Conformal-gated manifold adaptation further constrains updates during sustained OOD intervals to reduce contamination of the in-distribution reference.

As a neural sequential baseline, we evaluate a causal Temporal Transformer operating on the standardized novelty score stream $\{Z_t\}$. The model uses context window $T = 32$, $L = 2$ Transformer layers, $H = 4$ attention heads, and hidden dimension $D = 128$. It is trained on ID-only streams using one-step prediction loss under causal masking, and the prediction residual is used as the frame-level OOD score. "Temporal Transformer (ID-only)" denotes this predictive baseline trained only on ID streams, while "Temporal Transformer + Conformal" applies the same conformal-gated update rule as Eq. (8)–(9). STREAM-OOD corresponds to the full integration of multi-granularity representation modeling, nonparametric streaming manifold estimation, conformal-gated adaptation, and adaptive Bayesian regime inference.

References

[1] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. arXiv preprint arXiv:2209.09858, 2022. 5

[2] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136, 2016. 5

[3] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690, 2017. 5

[4] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21464–21475, 2020. 5

[5] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. arXiv preprint arXiv:2106.09022, 2021. 5

[6] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. arXiv preprint arXiv:2111.09805, 2021. 5

[7] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. arXiv preprint arXiv:2111.12797, 2021. 5

[8] Yiyou Sun, Yifei Ming, Xuhui Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 5

[9] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5