STREAMING ANOMALY DETECTION IN CONTAMINATED MANIFOLD-STRUCTURED HIGH-DIMENSIONAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional observations, such as material spectra and industrial process measurements, often concentrate near low-dimensional manifolds and arrive as streams during continuous scanning and process monitoring. Detecting anomalies in these streams requires learning emerging normal patterns without absorbing sparse anomalies into the background. Both emerging normal patterns and anomalies can receive high anomaly scores from a model trained on earlier observations, so its scores alone cannot determine which samples should guide subsequent learning. We propose Evolving Score Fields (ESF), a streaming anomaly detection framework for manifold-structured observations. By learning gradients of the log density, its score model implicitly represents the background distribution and its local geometry. ESF distinguishes compatibility with the learned background from support within the current block. Score-direction consistency measures departure from the learned field, while neighborhood support allows well-supported new patterns to enter adaptation even when the existing field assigns them high anomaly scores. Selected samples update the field through local denoising score matching with LoRA adapters and historical replay. The adapted field scores the current block, and a sample-aware merge carries the update into subsequent blocks. This connects anomaly detection and background learning within one evolving field, without anomaly labels or full-model retraining. Experiments on HyperAD and sparse-fault TEP streams demonstrate ESF's strong detection performance across spectral and industrial data.

1 INTRODUCTION

High-dimensional observations, including material spectra, industrial signals, and image representations, often concentrate near lower-dimensional manifolds (Bengio et al., 2013). This structure provides a basis for anomaly detection: normal observations occupy a small part of the ambient space, and departures from that structure can indicate anomalies (Pang et al., 2021; Lai et al., 2020). In a stream, however, normal conditions change with location, material composition, or operating state (Gama et al., 2014). A detector must learn these changes from incoming data that may also contain anomalies.

Existing streaming detectors track changing normality through compact online statistics, randomized partitions, or adaptation of domain-specific models (Guha et al., 2016; Manzoor et al., 2018; Kim et al., 2024; Garske et al., 2025). However, summaries based on means and covariances, as used in ERX (Garske et al., 2025), do not fully capture general nonlinear manifold geometry. As new normal components emerge, a detector must also learn their structure from observations that may contain anomalies. We therefore focus on streaming anomaly detection in high-dimensional data with manifold-structured backgrounds, seeking to combine geometric background modeling with online adaptation under contamination.

Neural score fields offer a flexible background representation for this purpose. A score-based generative model estimates the gradient of the log density, $s(x) = \nabla_x \log p(x)$, at different noise levels (Song & Ermon, 2019; Song et al., 2021). Learning these gradients yields an implicit model of the data distribution and its underlying manifold structure, without requiring an explicit density func-

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.

tion or manifold parameterization (Alain & Bengio, 2014; Pidstrigach, 2022). The learned field describes local directions toward higher background density. Around a locally smooth normal structure, perturbations on different sides induce score directions that tend to cancel; away from that structure, the directions acquire a shared component toward the background (Sheng et al., 2026). Their consistency thus provides a geometric basis for anomaly detection.

A useful background representation must also evolve as normality changes (Kim et al., 2024; 2026). A field learned from earlier observations can assign high anomaly scores both to genuine anomalies and to a previously unseen normal component. These cases require opposite learning decisions: the new background should enter the model, whereas anomalies should not. The field therefore faces a learning problem that its own anomaly scores cannot resolve. Rejecting high-score observations can exclude emerging normal patterns; fitting the entire stream can instead make anomalies part of the modeled background, a failure we call *anomaly absorption*. Online adaptation needs evidence of normality beyond agreement with the existing model.

We study streaming anomaly detection in contaminated, manifold-structured high-dimensional data. An unlabeled initial dataset $\mathcal{D}_0$, dominated by normal samples, provides a warm start, followed by unlabeled blocks $\mathcal{B}_1, \mathcal{B}_2, \ldots$ in which normal observations remain the majority. The background visible in a block changes as observation proceeds: some normal components disappear from view while others emerge. A component does not become anomalous simply because it is no longer observed. Figure 1 distinguishes the currently visible background from this accumulated normal reference. The detector learns emerging structure from incoming blocks and revisits earlier examples through finite replay.

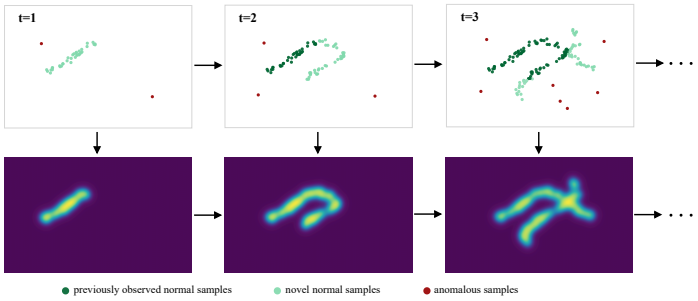

Figure 1: **Schematic illustration of normal-background expansion in a contaminated data stream.**

We propose Evolving Score Fields (ESF), a unified framework that uses an evolving neural score field for both anomaly detection and background learning. ESF distinguishes learned-background compatibility from current-block support: score-direction consistency identifies departures from the learned field, while neighborhood support measures evidence within the arriving block. When unfamiliar observations become widespread, selection shifts toward current-block support. Well-supported new components can therefore supply training data before the field recognizes them as normal. Updating the field changes the geometry used for anomaly scoring, connecting background discovery to detection.

ESF implements this learning process through local adaptation and consolidation. Local denoising score matching fits LoRA adapters to selected samples while keeping the base network frozen (Vincent, 2011; Hu et al., 2022); replay contributes earlier background examples (Rolnick et al., 2019). The adapted field scores the current block. A sample-aware merge incorporates part of the update into the model used for subsequent blocks, balancing newly selected data against historical effective sample mass. This formulation applies to vector-valued observations across domains, with dataset-specific preprocessing and training settings.

Our main experiments use the Synthetic, SNP, and Beach scenes from HyperAD (Garske et al., 2025) to evaluate detection as spectral backgrounds change during scanning. ESF achieves the highest AUPRC on all three scenes. To examine applicability beyond spectral data, we evaluate the same framework on TEP streams with sparse fault observations (Reinartz et al., 2021). These streams test 28 fault types along a shared change in normal operating conditions, where ESF achieves competitive results. Together, these experiments establish spectral anomaly detection as the primary application and demonstrate applicability to industrial process measurements.

Our contributions are summarized as follows:

- We develop a score-field framework that connects streaming anomaly detection with online background adaptation under contamination, using one evolving representation for both tasks.

- We combine learned-field compatibility with current-block support to admit emerging normal patterns before the field recognizes them. Local low-rank fitting and consolidation incorporate this evidence without full-model retraining.
- We evaluate the framework on three HyperAD scenes and demonstrate its applicability beyond spectral data on sparse-fault TEP streams.

2 Related Work

Streaming anomaly detection. Streaming anomaly detection identifies unusual observations as data arrive, often under changes in the normal distribution (Gama et al., 2014). Classical online methods maintain compact statistics, randomized partitions, or density surrogates, allowing efficient updates without retaining the full stream (Tan et al., 2011; Guha et al., 2016; Manzoor et al., 2018; Bhatia et al., 2021; Leveni et al., 2024). Neural approaches also support online adaptation: ARCUS maintains a pool of autoencoder-based detectors and updates the pool as the data distribution evolves (Yoon et al., 2022).

Industrial process monitoring is a major application, commonly studied as multivariate time-series anomaly detection. Deep detectors model normal operation through reconstruction (Audibert et al., 2020; Tuli et al., 2022), probabilistic sequence modeling (Su et al., 2019), association discrepancy (Xu et al., 2022), graph-based variable dependencies (Deng & Hooi, 2021), and frequency representations (Wu et al., 2025). Moving beyond fixed training, M2N2 learns new normal patterns during inference (Kim et al., 2024), while CANDI selects potential false positives for adaptation using anomaly scores and latent similarity (Kim et al., 2026). These methods highlight the need to distinguish changes in normal behavior from anomalies when learning from unlabeled observations.

Hyperspectral line scanning provides another streaming setting, where background materials can change as a push-broom sensor acquires successive image lines (Horstrand et al., 2019). RX measures spectral deviations using background mean and covariance estimates (Reed & Yu, 1990); causal variants update these estimates from arriving observations (Chen et al., 2014), and ERX uses exponential updates to track changing scenes (Garske et al., 2025). In ESF, current-block support guides adaptation of the same neural score field used for detection, admitting unfamiliar normal structure before it is represented by the learned geometry.

Score-based generative models. Score-based generative models learn a data distribution through its score, the gradient of the log density (Song & Ermon, 2019). Score matching estimates unnormalized models without evaluating their normalizing constants (Hyvärinen, 2005), while denoising score matching learns the score from noise-perturbed data (Vincent, 2011). For a perturbed sample x_τ with marginal density p_τ, a time-conditioned network approximates

$$s_\theta(x_\tau, \tau) \approx \nabla_{x_\tau} \log p_\tau(x_\tau), \tag{1}$$

where τ denotes perturbation time. In continuous-time generative modeling, these scores determine the reverse-time dynamics associated with the forward noising process (Song et al., 2021). Beyond generation, learning the score implicitly represents the noise-smoothed data distribution and its local geometry. For manifold-structured data, this provides a way to model background structure without explicitly parameterizing the manifold (Alain & Bengio, 2014; Pidstrigach, 2022). ESF uses this geometric information for anomaly detection and continually adapts the field to newly observed normal components.

Parameter-efficient adaptation. Parameter-efficient adaptation trains a small set of parameters while keeping most pretrained weights fixed, as in adapter tuning and prefix tuning (Houlsby et al., 2019; Li & Liang, 2021). Low-rank adaptation (LoRA) freezes an existing weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and parameterizes its update as (Hu et al., 2022)

$$W' = W + \Delta W, \qquad \Delta W = \frac{\alpha}{r} BA, \tag{2}$$

where $A \in \mathbb{R}^{r \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$, and $r \ll \min(d_{\text{in}}, d_{\text{out}})$. LoRA reduces the number of optimized parameters and allows the learned increments to be merged into the base weights. In ESF, each temporary adapter fits a local change to the score field, and sample-aware consolidation incorporates a fraction of that update into the background model. Repeating this process evolves a single score field without full-model retraining.

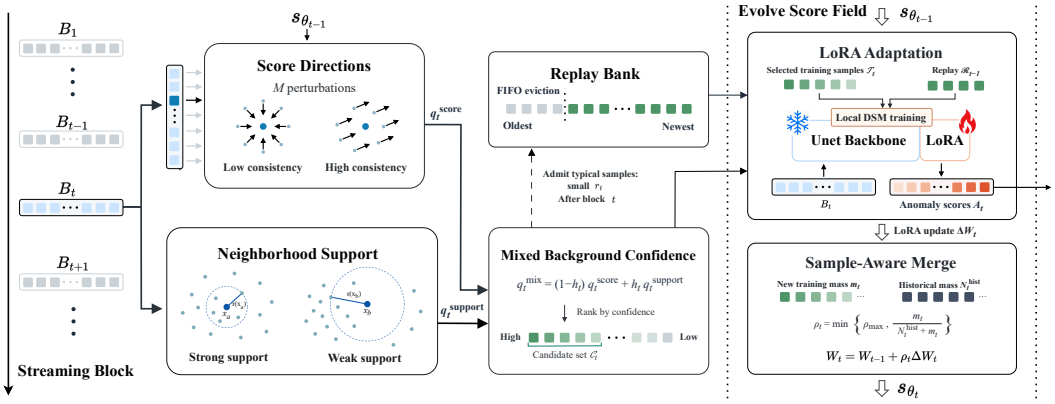

Figure 2: **Overview of ESF.** $\mathcal{B}_t$ denotes the current data block. We select samples using score-direction consistency and neighborhood support, then perform LoRA adaptation with historical replay. The adapted field then scores the current block. Finally, sample-aware merging carries the update forward to subsequent blocks.

3 METHOD

ESF uses one neural score field for streaming anomaly detection and background learning, as shown in Figure 2. Its update combines agreement with the learned field and support within the current block, allowing unfamiliar but well-supported patterns to enter adaptation. The adapted field scores the block, and consolidation carries the update forward.

3.1 PROBLEM FORMULATION

We observe unlabeled high-dimensional vectors $x \in \mathbb{R}^d$ in an initial dataset $\mathcal{D}_0 = \{x_i^{(0)}\}_{i=1}^{n_0}$, followed by streaming blocks $\mathcal{B}_t$ of size n_t. Normal samples dominate both initialization and the stream, while sparse anomalies may occur throughout.

Each arriving block follows a contaminated distribution

$$p_t(x) = (1 - \pi_t)p_t^{\mathrm{n}}(x) + \pi_t p_t^{\mathrm{a}}(x), \tag{3}$$

where p_t^{n} and p_t^{a} denote the current normal and anomaly distributions, respectively, and π_t is the minority anomaly fraction. Neither component labels nor π_t are observed.

Normal observations concentrate near lower-dimensional manifolds rather than lying exactly on them. Locally, we write

$$x = f_t(u) + \xi, \qquad u \sim p_t^{\mathrm{latent}}(u), \qquad \xi \sim \mathcal{N}(0, \sigma_{\mathrm{obs}}^2 I), \tag{4}$$

where $u \in \mathbb{R}^{d_{\mathcal{M}}}$, $d_{\mathcal{M}} < d$, parameterizes a normal component through f_t, and ξ represents observation noise. The latent density describes variation along the component, while noise spreads its probability mass into the surrounding region. Several such components may coexist within a block.

The observed background changes, but a normal component remains valid after it leaves the current view. We define the accumulated normal reference by

$$\mathcal{M}_{\le t}^{\mathrm{n}} = \bigcup_{j=0}^{t} \mathcal{M}_j^{\mathrm{n}}. \tag{5}$$

This union defines which structures count as normal; the learned field's coverage depends on adaptation and retained data. Let $p_{\le t}^{\mathrm{n}}$ denote a reference distribution over these components (Appendix B), which ESF aims to approximate through its score field.

At block t, the detector outputs $A_t(x)$ for every $x \in \mathcal{B}_t$ and updates its parameters from θ_{t-1} to θ_t. The score-field learning target is

$$s_{\theta_t}(x, \tau) \approx \nabla_x \log p_{\le t,\tau}^{\mathrm{n}}(x), \tag{6}$$

where τ is diffusion time and $p^{\mathrm{n}}_{\leq t,\tau}$ is the noise-smoothed accumulated background distribution. Incorporating anomalous structure into this background is *anomaly absorption*.

Adaptation uses only initial, current, and retained past observations, without labels. Scores are emitted after adapting to the current block.

3.2 Score Field and Background Evidence

Initial score field. The score field implicitly models the background distribution and its local geometry. We use an unconditional ResMLP U-Net for vector-valued observations; architectural details are given in Appendix B.

We learn the initial field by denoising score matching (DSM) (Vincent, 2011). Let $\sigma(\tau)$ denote the perturbation standard deviation at diffusion time τ; the variance-exploding (VE) schedule is specified in Appendix B.1. The per-sample objective is

$$\ell(s; x, \tau, z) = \|\sigma(\tau)s(x + \sigma(\tau)z, \tau) + z\|_2^2 . \tag{7}$$

The initial field minimizes

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_0,\, \tau\sim\mathcal{U}(\tau_{\min},1),\, z\sim\mathcal{N}(0,I)}[\ell(s_\theta; x, \tau, z)]. \tag{8}$$

This learns the score of a noise-smoothed distribution around the observed background structure (Song & Ermon, 2019; Song et al., 2021). The field therefore describes a neighborhood of that structure, rather than requiring normal observations to lie exactly on an idealized manifold.

Direction-consistency evidence. Around a low-dimensional background manifold, perturbations mainly explore directions normal to the manifold; their scores point back from different sides and tend to cancel. Far from the manifold relative to the perturbation radius, the scores retain a common inward direction. We measure this alignment through direction consistency. Let $u_s(y, \tau) = s(y, \tau)/\|s(y, \tau)\|_2$, with $u_s = 0$ for a zero score. At inference time $\tau^\star$,

$$A(x; s) = \left\| \frac{1}{M} \sum_{m=1}^{M} u_s(x + \sigma(\tau^\star)\epsilon_m, \tau^\star) \right\|_2 , \qquad \epsilon_m \sim \mathcal{N}(0, I). \tag{9}$$

Normalizing each score makes $A(x; s) \in [0, 1]$ measure angular agreement rather than magnitude, without reverse-diffusion sampling. Under a local Gaussian-tube reference model, population consistency and the expected squared anomaly score for any $M > 1$ both increase strictly with normal distance (Propositions A.1 and A.2).

The existing field gives provisional scores $\widetilde{A}_t(x) = A(x; s_{\theta_{t-1}})$. A large value indicates directional agreement, but can arise from either an anomaly or a newly emerging normal component. Selecting training samples solely by this score would risk excluding the very background components that the model needs to learn.

Neighborhood support evidence. A new normal component can be unfamiliar to the learned field yet well supported within the current block. We measure support by the radius $r_t(x)$ to the K_t-th nearest other sample, with $K_t = \min\{\lceil \kappa n_t \rceil, n_t - 1\}$; smaller radii indicate stronger support. Choosing K_t above the anticipated anomaly count requires an isolated anomaly group to reach beyond its own members. Distances use fixed initial-data scaling (Appendix B.2).

Let $\mathcal{A}_t = \{\widetilde{A}_t(x) : x \in \mathcal{B}_t\}$ and $\mathcal{D}_t = \{r_t(x) : x \in \mathcal{B}_t\}$. We convert score-field compatibility and neighborhood support into background confidences:

$$\begin{aligned} q_t^{\mathrm{score}}(x) &= \exp\left(-\frac{[\widetilde{A}_t(x) - Q_{\alpha_s}(\mathcal{A}_t)]_+}{c_t^s} \right), \\ q_t^{\mathrm{support}}(x) &= \exp\left(-\frac{[r_t(x) - Q_{\alpha_r}(\mathcal{D}_t)]_+}{c_t^r} \right). \end{aligned} \tag{10}$$

Here $[v]_+ = \max\{v, 0\}$, Q_α is an empirical quantile, and c_t^s, c_t^r are the respective IQRs with positive floors (Appendix B.2). Each confidence equals one below its block-specific quantile and decreases exponentially above it.

When high provisional scores become widespread, sample selection places more weight on current-block support. We measure this change relative to the initial score reference $T_0 = Q_{\alpha_0}(\mathcal{A}_0)$, where $\mathcal{A}_0 = \{A(x; s_{\theta_0}) : x \in \mathcal{D}_0\}$ and $0 < \alpha_0 < 1$:

$$a_t = \frac{1}{n_t} \sum_{x \in \mathcal{B}_t} \mathbf{1}[\widetilde{A}_t(x) > T_0], \qquad e_t = \left[\frac{a_t - (1 - \alpha_0)}{\alpha_0} \right]_+ . \tag{11}$$

The excess fraction $e_t \in [0, 1]$ controls a state that persists as the field adapts:

$$h_t = \max\{e_t, (1 - \rho_{t-1})h_{t-1}\}, \qquad q_t^{\text{mix}} = (1 - h_t)q_t^{\text{score}} + h_t q_t^{\text{support}}, \tag{12}$$

Here $h_0 = \rho_0 = 0$ and ρ_{t-1} is the previous merge coefficient. As unfamiliar observations become widespread, h_t shifts selection toward neighborhood support; it decays as updates are consolidated. Retaining this response during adaptation allows a new normal component to supply training samples before the learned field fully represents it.

3.3 Confidence-Gated Low-Rank Adaptation

Selection-based gating. The top $\lfloor f_c n_t \rfloor$ samples under q_t^{mix} form $\mathcal{C}_t$, split into a training set $\mathcal{T}_t$ and a disjoint probe set $\mathcal{V}_t$. Gating acts through selection: admitted training samples receive equal weight. Both sets remain fixed during optimization, and every observation in $\mathcal{B}_t$ is scored regardless of selection.

Low-rank fitting. We represent block-local adaptation of the score field with a temporary LoRA adapter (Hu et al., 2022). For each adapted content-transform layer ℓ,

$$\Delta W_{t,\ell} = \frac{\alpha_{\text{L}}}{r} B_{t,\ell} A_{t,\ell}, \qquad W_{t,\ell}^{\text{fit}} = W_{t-1,\ell} + \Delta W_{t,\ell}. \tag{13}$$

We denote these increments collectively by Δ_t and optimize only the LoRA factors, with the base model frozen. Each block starts with zero adapter increment.

Online fitting uses a time distribution P_{loc} concentrated at and around the detection scale $\tau^\star$ (Appendix B.1). Aligning fitting with detection focuses each block's update on the local score geometry used to assess anomalies. Its local DSM loss is

$$\bar{\ell}_t(x; \Delta) = \mathbb{E}_{\tau \sim P_{\text{loc}},\, z \sim \mathcal{N}(0,I)}[\ell(s_{\theta_{t-1},\Delta}; x, \tau, z)]. \tag{14}$$

Averaging over current samples and the replay bank gives

$$\mathcal{L}_{\text{current}} = \frac{1}{|\mathcal{T}_t|} \sum_{x \in \mathcal{T}_t} \bar{\ell}_t(x; \Delta), \qquad \mathcal{L}_{\text{replay}} = \frac{1}{|\mathcal{R}_{t-1}|} \sum_{x \in \mathcal{R}_{t-1}} \bar{\ell}_t(x; \Delta). \tag{15}$$

The adaptation objective is

$$\mathcal{L}_t = (1 - \beta_t)\mathcal{L}_{\text{current}} + \beta_t \mathcal{L}_{\text{replay}} + \lambda_{\text{reg}} \mathcal{R}_{\text{LoRA}}, \qquad \beta_t = \beta_0(1 - h_t), \tag{16}$$

where $\mathcal{R}_{\text{LoRA}} = \sum_\ell (\|A_{t,\ell}\|_F^2 + \|B_{t,\ell}\|_F^2)$. Current samples extend the field toward newly observed normal structure, while replay revisits earlier backgrounds that may no longer appear in the stream. As h_t increases, β_t decreases, giving the arriving background greater influence during adaptation.

We retain the adapter with the lowest checked local DSM loss on $\mathcal{V}_t$, including the zero-update state. Fixed probe perturbations allow consistent comparisons across updates; the disjoint probes assess the fit beyond the samples supplying gradients. Optimization controls are given in Appendix B.4.

3.4 Score-Field Consolidation and Replay

Sample-aware consolidation. A local adapter reflects the observations in one block. Consolidation determines how much of that change enters the background model used by subsequent blocks. We measure new evidence by distinct selected training samples, rather than repeated optimization steps, and balance their count $m_t = |\mathcal{T}_t|$ against historical effective mass N_{t-1}:

$$N_t^{\text{hist}} = \eta N_{t-1}, \qquad \rho_t = \min\left\{ \rho_{\max}, \frac{m_t}{N_t^{\text{hist}} + m_t} \right\}, \tag{17}$$

with $N_t = \min\{N_{\max}, N_t^{\text{hist}} + m_t\}$. Decay $0 < \eta < 1$ and capping prevent accumulated mass from suppressing later updates; $\rho_{\max}$ limits each merge. Initial mass is defined in Appendix B.3.

We then compute the final anomaly scores for all samples in $\mathcal{B}_t$ by applying the direction-consistency measure in Eq. (9) to the adapted score field:

$$\gamma_t = \rho_t + (1 - \rho_t)h_t, \qquad A_t(x) = A\big(x; s_{\theta_{t-1}, \gamma_t \Delta_t}\big)\,. \tag{18}$$

Thus, $\gamma_t \in [\rho_t, 1]$ gives the local fit greater influence when unfamiliar observations are widespread. We then consolidate the base weights,

$$W_{t,\ell} = W_{t-1,\ell} + \rho_t \Delta W_{t,\ell}, \tag{19}$$

and reset the adapters. The next adapter is fitted on top of this updated model. Local changes thus accumulate in a shared score field, rather than forming a growing collection of block-specific models. This connects adaptation to the current background with background model updates across successive blocks.

Typical-background replay. After scoring and consolidation, we admit the most strongly supported candidates from $\mathcal{C}_t$ to a bounded FIFO replay bank (Rolnick et al., 2019), at rate f_r relative to the full block. They provide historical examples from the next block onward, maintaining a training signal for earlier normal structure as the field expands. Algorithm 1 and Appendix B.2 give the full procedure and sampling details.

4 Experiments

4.1 Experimental setup

Datasets. Our main evaluation uses three HyperAD scenes: Synthetic, SNP, and Beach (Garske et al., 2025), which test adaptation to changing spectral backgrounds. We then assess applicability to industrial observations using sparse-fault streams constructed from the Tennessee Eastman Process (TEP) (Reinartz et al., 2021). Qualitative vessel and fire-point applications appear in Appendix E.

Baselines. For spectral imagery, we use the HyperAD implementations of RX (Reed & Yu, 1990), RT-CK-RXD (Chen et al., 2014), RX-BIL (Du & Nekovei, 2009), LBL-AD (Horstrand et al., 2019), and ERX (Garske et al., 2025). On TEP, we compare Window-kNN, Online-iForest (Leveni et al., 2024), ERX, xStream (Manzoor et al., 2018), and ARCUS (Yoon et al., 2022) with RSRAE (Lai et al., 2020) and RaPP (Kim et al., 2020) backbones.

Evaluation. We report AUROC and AUPRC (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015), causal F1, and F1. Causal F1 uses a past-score threshold at quantile $q = 0.99$; F1 uses the best single fixed threshold selected with evaluation labels and serves as an offline reference. Both benchmarks additionally use AUC-TD and AUC-BS for target detectability and background suppression (Chang, 2021). Methods share the post-initialization evaluation samples; labels do not enter adaptation or causal thresholding. Metric definitions and implementation settings are given in Appendices C and B, with TEP details in Appendix C.4.

4.2 Streaming hyperspectral anomaly detection

HyperAD images are processed column by column to simulate real-world airborne spectral scanning. The first 99 columns initialize the detector, and the remaining columns form the evaluation stream. ESF divides spectra by the maximum of the initial columns and holds this scale fixed without clipping later observations. Baselines retain their native spectral inputs. Model updates use only the current and earlier blocks. Labels are used only for metric evaluation.

Synthetic and Beach contain hyperspectral observations, while SNP is multispectral. Table 1 summarizes scene-level detection, and Figure 3 presents the HyperAD scene previews, ground truth, and detection maps from ERX and ESF. Maps for all baselines are provided in Appendix D.4.

Across three runs, ESF obtains the highest mean AUPRC and AUC-TD on all three scenes. Its mean AUPRC is 0.4589 on Synthetic, 0.8749 on SNP, and 0.4111 on Beach, compared with 0.0507,

0.5330, and 0.2187 for the strongest baseline on each scene. It also leads mean AUROC, causal F1, and F1 on every scene. These results demonstrate ESF's strong detection performance across the three scenes.

In Figure 3, target responses are generally more visually prominent with ESF than with ERX across the three scenes. The SNP and Beach maps exhibit localized target responses.

Table 1: **Streaming hyperspectral anomaly detection.** Three-seed means on post-initialization regions. Best: **bold**; second: underlined.

Dataset	Method	AUROC (↑)	AUPRC (↑)	Causal F1 (↑)	F1 (↑)	AUC-TD (↑)	AUC-BS (↑)
Synthetic	RX	0.6442	0.0507	0.0682	0.0961	0.8255	0.5000
	RT-CK-RXD	0.5816	0.0358	0.0549	0.0684	0.7569	0.4331
	RX-BIL	0.5665	0.0395	0.0573	0.0704	0.7467	0.4114
	LBL-AD	0.5725	0.0151	0.0028	0.0494	1.1610	-0.0148
	ERX	0.7776	0.0348	0.0443	0.0710	1.0334	0.5867
	ESF	**0.8805**	**0.4589**	**0.1398**	**0.4858**	**1.2760**	**0.7195**
SNP	RX	0.9886	0.2274	0.1254	0.3167	1.2307	**0.9104**
	RT-CK-RXD	0.9918	0.2840	0.1269	0.3979	1.2754	0.9036
	RX-BIL	0.9917	0.2723	0.1259	0.3861	1.2788	0.8949
	LBL-AD	0.9705	0.5330	0.1299	0.6040	1.4607	0.8261
	ERX	0.9729	0.2433	0.1210	0.3491	1.2607	0.8736
	ESF	**0.9933**	**0.8749**	**0.1310**	**0.8282**	**1.9223**	0.6905
Beach	RX	0.8288	0.1727	0.0611	0.3438	1.0278	0.7138
	RT-CK-RXD	0.8057	0.2187	0.0669	0.4196	1.0202	0.6821
	RX-BIL	0.8031	0.2183	0.0665	0.4191	1.0128	0.6834
	LBL-AD	0.6970	0.0768	0.0700	0.2080	0.8332	0.5988
	ERX	0.9387	0.1739	0.0443	0.2963	1.1303	0.8334
	ESF	**0.9527**	**0.4111**	**0.1209**	**0.5314**	**1.2927**	**0.8401**

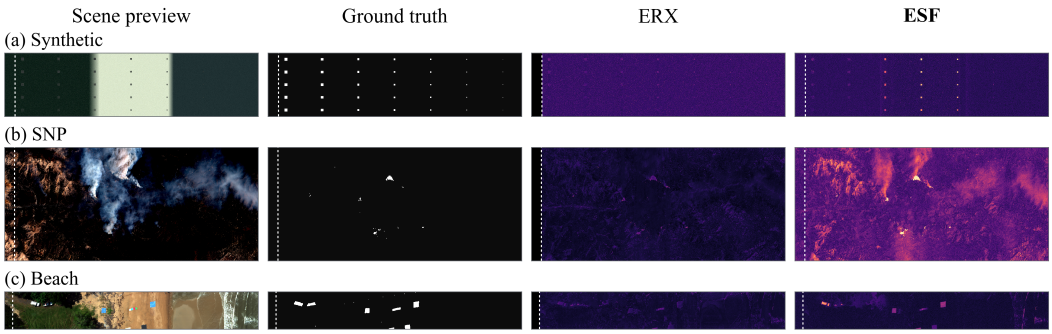

Figure 3: **HyperAD detection maps.** Scene previews, ground truth, and detection maps from ERX and ESF. White pixels in the ground-truth maps indicate anomalies. The first 99 columns, marked by dashed lines, form the initialization region and are excluded from evaluation. Images are scanned column by column from left to right.

4.3 Industrial process anomaly detection

On TEP, 28 sparse-fault streams share an initialized normal Mode-1-to-Mode-2 trajectory (Reinartz et al., 2021). Each replaces 74 observations with mode-matched fault samples; the transition remains normal.

All methods receive the same initialization-fitted transform of the 85 process variables. Table 2 reports six metrics averaged over the 28 post-initialization fault streams for each of three runs; Appendix C.4 gives the construction and evaluation details. Run-to-run variation for both benchmarks appears in Appendix D.1.

ESF leads mean AUROC, AUPRC, both F1 measures, and AUC-TD across the 28 fault streams. Window-kNN leads AUC-BS.

Its mean AUPRC reaches 0.7788 versus Window-kNN's 0.6430, while mean causal F1 reaches 0.3223 versus ARCUS-RSRAE's 0.2599. The same score model and adaptation procedure apply to these industrial process variables.

Table 2: **Anomaly detection on TEP.** Three-seed means over 28 fault types.

Method	AUROC	AUPRC	Causal F1	F1	AUC-TD	AUC-BS
Window-kNN	0.8968	<u>0.6430</u>	0.1916	<u>0.6402</u>	1.0867	**0.8649**
Online-iForest	0.5943	0.0992	0.1387	0.1612	0.8079	0.4203
ERX	0.5589	0.1154	0.0609	0.1425	1.1297	0.0395
xStream	0.6674	0.0856	0.0841	0.1955	1.3312	0.1929
ARCUS-RSRAE	0.5621	0.3567	<u>0.2599</u>	0.4124	1.0508	0.1346
ARCUS-RaPP	<u>0.9006</u>	0.4757	0.2269	0.5778	<u>1.6257</u>	0.3792
ESF	**0.9496**	**0.7788**	**0.3223**	**0.8054**	**1.7308**	<u>0.7940</u>

4.4 Ablation studies

Table 3 isolates online updates, confidence selection, neighborhood support, and replay on the HyperAD streams. Appendix C.5 defines the variants.

Online adaptation raises mean AUROC from 0.7451 to 0.9413. Yet random candidate selection lowers AUPRC to 0.1700, below the frozen field's 0.4200. The choice of training samples is therefore central to learning from contaminated blocks.

Table 3: **Component ablations on HyperAD.** Three-scene means with seed 1.

Variant	AUROC	AUPRC	Causal F1	F1	AUC-TD	AUC-BS
w/o online update	0.7451	0.4200	0.1029	0.4555	1.2394	0.4618
w/o conf. selection	0.8152	0.1700	0.0716	0.2707	1.1348	0.6045
w/o q^{support}	0.9100	0.4689	0.1118	0.5437	1.4296	0.7075
w/o replay	<u>0.9399</u>	<u>0.5619</u>	<u>0.1277</u>	**0.6187**	<u>1.4887</u>	**0.7521**
ESF (full)	**0.9413**	**0.5630**	**0.1279**	<u>0.6171</u>	**1.4985**	<u>0.7473</u>

Using only score-field confidence for selection lowers AUPRC from 0.5630 to 0.4689, showing the value of neighborhood support when unfamiliar normal spectra enter a block. In the controlled A-to-B stream of Appendix D.2, B enters training as it becomes prevalent and its held-out score falls. Under persistent 20% contamination, ESF admits 3 anomalies among 1,224 selected training samples, versus 230 under equal-size random selection. Replay is designed to retain earlier background structure so that returning normal patterns are not mistaken for anomalies. Such returns are not prominent in HyperAD, and removing replay changes mean AUPRC only from 0.5630 to 0.5619; the no-replay variant narrowly leads F1 and AUC-BS.

5 Conclusion

We presented Evolving Score Fields (ESF), a unified framework for streaming anomaly detection and online background adaptation in contaminated, manifold-structured data. ESF combines compatibility with the learned field and neighborhood support in the current block, allowing well-supported new normal patterns to guide adaptation before the model recognizes them as normal. Local adaptation and consolidation evolve the same score field used for anomaly scoring. Results on HyperAD and sparse-fault TEP streams demonstrate the effectiveness of this shared score-field framework across spectral and industrial data. Qualitative results on maritime vessel and forest fire-point detection further illustrate its use in remote-sensing applications.

AI use statement

Generative AI tools assisted with discussions of the method and experimental design, software implementation, synthetic-data generation, theoretical analysis, interpretation of experimental results, and drafting and translation of the manuscript. The authors reviewed the AI-assisted material and take responsibility for the paper's claims, results, and final text.

REFERENCES

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research*, 15(110):3743–3773, 2014.

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 International Conference on Knowledge Discovery and Data Mining*, pp. 3395–3404, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Siddharth Bhatia, Arjit Jain, Pan Li, Ritesh Kumar, and Bryan Hooi. MStream: Fast anomaly detection in multi-aspect streams. In *Proceedings of the Web Conference 2021*, pp. 3371–3382, 2021.

Chein-I Chang. An effective evaluation tool for hyperspectral target detection: 3D receiver operating characteristic curve analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5131–5153, 2021.

Shih-Yu Chen, Yulei Wang, Chao-Cheng Wu, Chunhong Liu, and Chein-I Chang. Real-time causal processing of anomaly detection for hyperspectral imagery. *IEEE Transactions on Aerospace and Electronic Systems*, 50(2):1511–1534, 2014.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, 2006.

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.

Qian Du and Reza Nekovei. Fast real-time onboard processing of hyperspectral imagery for detection and classification. *Journal of Real-Time Image Processing*, 4(3):273–286, 2009.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

Samuel Garske, Bradley Evans, Christopher Artlett, and K. C. Wong. ERX: A fast real-time anomaly detection algorithm for hyperspectral line scanning. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025.

Sudipto Guha, Nina Mishra, Gourav Roy, and Okke Schrijvers. Robust random cut forest based anomaly detection on streams. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2712–2721, 2016.

Pablo Horstrand, María Díaz, Raúl Guerra, Sebastián López, and José Fco. López. A novel hyperspectral anomaly detection algorithm for real-time applications with push-broom sensors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12):4787–4797, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Dongmin Kim, Sunghyun Park, and Jaegul Choo. When model meets new normals: Test-time adaptation for unsupervised time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13113–13121, 2024.

HyunGi Kim, Jisoo Mok, Hyungyu Lee, Juhyeon Shin, and Sungroh Yoon. CANDI: Curated test-time adaptation for multivariate time-series anomaly detection under distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15018–15026, 2026.

Ki Hyun Kim, Sangwoo Shim, Yongsub Lim, Jongseob Jeon, Jeongwoo Choi, Byungchan Kim, and Andre S. Yoon. RaPP: Novelty detection with reconstruction along projection pathway. In *International Conference on Learning Representations*, 2020.

Chieh-Hsin Lai, Dongmian Zou, and Gilad Lerman. Robust subspace recovery layer for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2020.

Filippo Leveni, Guilherme Weigert Cassales, Bernhard Pfahringer, Albert Bifet, and Giacomo Boracchi. Online isolation forest. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27288–27298, 2024.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Emaad Manzoor, Hemank Lamba, and Leman Akoglu. xStream: Outlier detection in feature-evolving data streams. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1963–1972, 2018.

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.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35852–35865, 2022.

I. S. Reed and X. Yu. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(10): 1760–1770, 1990.

Christopher Reinartz, Murat Kulahci, and Ole Ravn. An extended Tennessee Eastman simulation dataset for fault-detection and decision support systems. *Computers & Chemical Engineering*, 149:107281, 2021.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

scikit-learn developers. average_precision_score. scikit-learn API documentation, n.d. URL `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html`. Accessed September 19, 2026.

Jiahui Sheng, Yidan Shi, Shu Xiang, Xiaorun Li, and Shuhan Chen. Utilizing the score of data distribution for hyperspectral anomaly detection, 2026. arXiv preprint arXiv:2601.12379.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

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.

Swee Chuan Tan, Kai Ming Ting, and Tony Fei Liu. Fast anomaly detection for streaming data. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 1511–1516, 2011.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

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.

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.

Susik Yoon, Youngjun Lee, Jae-Gil Lee, and Byung Suk Lee. Adaptive model pooling for online deep anomaly detection from a complex evolving data stream. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2347–2357, 2022.

A SCORE-DIRECTION GEOMETRY

Score-direction consistency measures how coherently the local field points toward the background. We explain this geometry in an affine reference model: population consistency increases with normal distance, and the same dependence holds for the expected squared anomaly score with finitely many perturbations. We then examine how observation noise changes these distances.

A.1 DISTANCE AND DIRECTION CONSISTENCY

Geometric intuition. In a locally flat reference field, score vectors point back toward a background manifold. For a low-dimensional manifold, isotropic perturbations mainly explore normal directions. At an on-manifold point, symmetric perturbations produce opposite return directions, which cancel in expectation. Away from the manifold, the perturbed points share a normal displacement. When this displacement dominates the perturbation scale, their directions back to the manifold remain similar. The following model formalizes this transition from cancellation to alignment.

Local reference field. Let $\mathcal{M} \subset \mathbb{R}^d$ be an affine subspace of dimension $d_{\mathcal{M}} < d$, with codimension $k = d - d_{\mathcal{M}} \geq 1$. Decompose a point into tangential coordinates u and normal displacement $v = x - \Pi_{\mathcal{M}}(x)$. Consider the Gaussian density of the normal coordinates,

$$\phi_\nu(v) = (2\pi\nu^2)^{-k/2} \exp\left(-\frac{\|v\|_2^2}{2\nu^2}\right), \qquad \nu > 0. \tag{20}$$

A local product model $p_\parallel(u)\phi_\nu(v)$ has purely normal score wherever the tangential density $p_\parallel$ is constant. We use the associated ideal vector field

$$s_\nu^0(x) = -\frac{x - \Pi_{\mathcal{M}}(x)}{\nu^2}. \tag{21}$$

We extend this reference field algebraically to isolate the normal geometry, without assuming a uniform density on an unbounded manifold. Observation noise and diffusion smoothing contribute to ν, which changes the field magnitude but not its direction.

Use $\sigma_\star = \sigma(\tau^\star) > 0$ for the inference perturbation standard deviation, keeping $\tau^\star$ as diffusion time. For $v = x - \Pi_{\mathcal{M}}(x)$, let $a = \|v\|_2/\sigma_\star$ and draw $\widetilde{x} = x + \sigma_\star\epsilon$, $\epsilon \sim \mathcal{N}(0, I_d)$. Define

$$u^0(\widetilde{x}) = \frac{s_\nu^0(\widetilde{x})}{\|s_\nu^0(\widetilde{x})\|_2} \tag{22}$$

when the denominator is nonzero, with zero assigned otherwise. The zero case has probability zero under the perturbation distribution.

Proposition A.1 (Monotonicity of direction consistency)**.** *For the affine reference field in Equation 21,*

$$\left\|\mathbb{E}_\epsilon[u^0(x + \sigma_\star\epsilon)]\right\|_2 = h_k\left(\frac{\operatorname{dist}(x, \mathcal{M})}{\sigma_\star}\right), \tag{23}$$

where

$$h_k(a) = \mathbb{E}_{g\sim\mathcal{N}(0,I_k)}\left[\frac{a + g_1}{\sqrt{(a + g_1)^2 + \sum_{j=2}^k g_j^2}}\right]. \tag{24}$$

The function $h_k : [0, \infty) \to [0, 1)$ is strictly increasing, satisfies $h_k(0) = 0$, and approaches one as $a \to \infty$.

Proof. Only the normal projection of ϵ affects the reference field. Rotate normal coordinates so that $v = \sigma_\star a e_1$. For $g \sim \mathcal{N}(0, I_k)$, the normalized score in these coordinates is

$$u^0(x + \sigma_\star\epsilon) = -\frac{ae_1 + g}{\|ae_1 + g\|_2}. \tag{25}$$

Symmetry cancels the expected components orthogonal to e_1. The component along the inward direction $-e_1$ is precisely $h_k(a)$, so $\mathbb{E}u^0 = -h_k(a)e_1$.

To see why alignment increases, fix g and write $R^2 = \sum_{j=2}^{k} g_j^2$. For $k \geq 2$, $R > 0$ almost surely, and the inward component satisfies

$$\frac{\partial}{\partial a} \frac{a + g_1}{\sqrt{(a+g_1)^2 + R^2}} = \frac{R^2}{\left((a+g_1)^2 + R^2\right)^{3/2}} > 0. \tag{26}$$

Thus increasing a strictly increases this component for almost every fixed perturbation, and therefore also its expectation. For $k = 1$, $h_1(a) = 2\Phi(a) - 1$, where Φ is the standard normal cumulative distribution function, which is also strictly increasing.

At $a = 0$, Gaussian symmetry gives $h_k(0) = 0$; hence $h_k(a) > 0$ for $a > 0$, establishing Equation 23. As $a \to \infty$, the inward component tends to one almost surely and is bounded in absolute value by one, so dominated convergence gives $h_k(a) \to 1$. At every finite a, Gaussian perturbations have positive probability of not pointing exactly inward, giving $h_k(a) < 1$. □

Direction alignment is controlled by normal displacement relative to the perturbation scale. The normal perturbation has root-mean-square length $\sigma_\star\sqrt{k}$; when the displacement dominates this scale, the score directions become strongly aligned. For fixed codimension k, this transition is described by $h_k(\mathrm{dist}(x, \mathcal{M})/\sigma_\star)$, independent of the density width ν.

The population quantity in Equation 23 averages directions before taking their norm. With finitely many perturbations, cancellation leaves a residual. The next result characterizes this residual and its relation to normal distance exactly.

A.2 Finite perturbations

Fix x and draw M independent perturbations. Let $U_m = u^0(x + \sigma_\star \epsilon_m)$, $\mu = \mathbb{E}U_m$, and $A_M^0(x) = \|M^{-1}\sum_m U_m\|_2$.

Proposition A.2 (Finite-perturbation geometry)**.** *For the affine reference field in Equation 21 and any integer $M \geq 1$,*

$$\mathbb{E}\left[(A_M^0(x))^2\right] = \frac{1}{M} + \left(1 - \frac{1}{M}\right) h_k\left(\frac{\mathrm{dist}(x, \mathcal{M})}{\sigma_\star}\right)^2. \tag{27}$$

For $M > 1$, the expected squared anomaly score is strictly increasing with normal distance.

Proof. The directions have unit norm almost surely. Independence gives $\mathbb{E}\langle U_m, U_n\rangle = \|\mu\|_2^2$ for $m \neq n$, so

$$\begin{aligned}\mathbb{E}\left[(A_M^0(x))^2\right] &= \frac{1}{M^2}\left(M + \sum_{m\neq n} \mathbb{E}\langle U_m, U_n\rangle\right) \\ &= \frac{1}{M} + \left(1 - \frac{1}{M}\right)\|\mu\|_2^2.\end{aligned}$$

Proposition A.1 identifies $\|\mu\|_2$ with $h_k(\mathrm{dist}(x, \mathcal{M})/\sigma_\star)$. Since h_k is nonnegative and strictly increasing, its square is strictly increasing as well. □

On $\mathcal{M}$, the expected squared anomaly score equals $1/M$, and Jensen's inequality gives $\mathbb{E}[A_M^0(x)] \leq M^{-1/2}$. Away from $\mathcal{M}$, the distance-dependent term increases. Finite sampling therefore introduces a background residual while preserving the geometric ordering in expected squared anomaly score.

The same independence calculation gives the mean-squared error of the average direction:

$$\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M} U_m - \mu\right\|_2^2 = \frac{1 - \|\mu\|_2^2}{M}. \tag{28}$$

More perturbations reduce this error at a linear computational cost. The experiments use $M = 16$ to balance detection performance and computation.

A.3 Observation noise

Observation noise moves normal samples into a neighborhood of the manifold. Under the affine reference, write $x = y + \xi$ as in Equation 4, and let $P_{\mathcal{N}}$ project onto the normal coordinates. For an observation with $\|P_{\mathcal{N}}\xi\|_2 \le b$, the reverse triangle inequality gives

$$\big|\mathrm{dist}(x, \mathcal{M}) - \mathrm{dist}(y, \mathcal{M})\big| \le \|P_{\mathcal{N}}\xi\|_2 \le b. \tag{29}$$

Thus a clean normal point on $\mathcal{M}$ remains within distance b, while a clean anomaly at distance at least Δ remains at distance at least $\Delta - b$. When $\Delta > 2b$, these distance ranges are disjoint. Propositions A.1 and A.2 then give lower population consistency and, for $M > 1$, lower expected squared anomaly scores under the reference field to these noisy normal observations than to the noisy anomalies. The expectations are over inference perturbations at each fixed observation.

For the isotropic Gaussian noise in Equation 4, the root-mean-square normal displacement is $\sigma_{\mathrm{obs}}\sqrt{k}$. Larger noise broadens the normal region and reduces its distance separation from anomalies. The scoring mechanism therefore accommodates observations near a manifold, with distinguishability governed by normal displacement relative to the inference perturbation scale.

Together, these results explain cancellation near a represented background component and alignment farther away. The affine reference isolates this local mechanism; curvature, tangential density variation, and score-model error can change the directions of a learned field.

B Additional Method Details

Accumulated background. The reference distribution in Section 3.1 expresses the relative emphasis on normal components observed over time:

$$p^{\mathrm{n}}_{\le t}(x) = \sum_{j=0}^{t} \omega_{j,t} p^{\mathrm{n}}_j(x), \qquad \omega_{j,t} > 0, \quad \sum_{j=0}^{t} \omega_{j,t} = 1. \tag{30}$$

The weights describe the learning target rather than explicit parameters maintained by ESF. Gaussian smoothing at scale $\sigma(\tau)$ gives $p^{\mathrm{n}}_{\le t,\tau}$, whose score defines the target in Equation 6. Appendix A analyzes score directions near a represented component of this background.

Algorithm 1 specifies the blockwise procedure. The following settings apply to HyperAD; TEP settings appear in Appendix C.4.

Algorithm 1 ESF Streaming Adaptation

Train s_{θ_0} on $\mathcal{D}_0$ using global DSM in Eq. (8).
Initialize the score and feature references, N_0, and $\mathcal{R}_0$; set $h_0 = \rho_0 = 0$.
for block $\mathcal{B}_t$, $t = 1, 2, \ldots$ **do**
  Score $\mathcal{B}_t$ with $s_{\theta_{t-1}}$ and compute its neighborhood radii.
  Compute h_t and q_t^{mix}; select $\mathcal{C}_t$ and split it into $\mathcal{T}_t, \mathcal{V}_t$.
  Fit reset LoRA factors using Eq. (16) and historical $\mathcal{R}_{t-1}$.
  Restore the best probe-checked adapter Δ_t.
  Compute m_t, N_t, ρ_t and γ_t; output A_t using Eq. (18).
  Merge $\rho_t \Delta_t$ into θ_{t-1} to obtain θ_t; reset the adapters.
  Admit typical samples from $\mathcal{C}_t$ and apply FIFO eviction to obtain $\mathcal{R}_t$.
end for

B.1 Model and noise

The unconditional ResMLP U-Net has two encoder stages, a bottleneck, and two decoder stages with concatenated skip connections, using LayerNorm, SiLU, and twofold MLP expansion without dropout. Its output matches the input dimension and is scaled by $1/\sigma(\tau)$. LoRA adapts 18 content-transform linear layers: input/output, both content layers per residual block, down/up projections, and skip fusion. All base-model parameters remain frozen during adapter optimization. Consolidation updates only the targeted base weight matrices; biases, normalization parameters, and time embeddings/projections remain unchanged.

Table 4: **Core ESF settings for HyperAD.**

Model / initialization	Value	Online adaptation	Value
Encoder / bottleneck	256, 128 / 64	LoRA rank / scale	8 / 16
Time embedding	256	Learning rate	3×10^{-3}
Residual blocks	5	Update steps	20–150
VE parameter b	5	Quantiles α_s, α_r	0.4, 0.75
Minimum time $\tau_{\min}$	10^{-5}	Reference quantile α_0	0.95
Inference time $\tau^\star$	0.05	Neighbor fraction κ	0.30
Perturbations M	16	Fractions f_c, f_{fit}	0.50, 0.80
Exact-time rate $p_\star$	0.5	Mass decay η	0.95
Noise ratios $a_{\min}, a_{\max}$	0.8, 1.25	Caps $\rho_{\max}, N_{\max}$	$0.5, 10^5$
Epochs	10	Replay weight β_0	0.05
Batch size	256	Replay fraction f_r	0.10
Sample cap	50,000	Replay capacity	4096
Learning rate	3×10^{-4}	Replay batch	64
Weight decay (both stages)	0.01	Factor penalty λ_{reg}	10^{-6}
Fourier scale	10	Gate scales c_s, c_r	$0.02, 10^{-6}$

We use $M = 16$ perturbations per sample to balance detection performance and computational cost. The variance-exploding perturbation schedule is

$$\sigma(\tau) = \sqrt{\frac{b^{2\tau} - 1}{2 \log b}}, \qquad b > 1, \tag{31}$$

with b specified in Table 4. The two stages sample noise as follows:

- **Initialization:** sample x uniformly from $\mathcal{D}_0$ and $\tau \sim \mathcal{U}[\tau_{\min}, 1]$.
- **Online fitting, replay, and probes:** choose $\tau^\star$ with probability $p_\star$; otherwise draw $\log a \sim \mathcal{U}(\log a_{\min}, \log a_{\max})$ and set $\tau = \sigma^{-1}(a\sigma(\tau^\star))$.

Here $\sigma^{-1}(v) = \log(1 + 2v^2 \log b)/(2 \log b)$, with times clipped to $[\tau_{\min}, 1]$.

B.2 Selection and replay

Selection and replay use the settings in Table 4. The neighborhood size uses a nominal anomaly-mass bound of 0.25.

The confidence scales are $c_t^s = \max\{\text{IQR}(\mathcal{A}_t), c_s\}$ and $c_t^r = \max\{\text{IQR}(\mathcal{D}_t), c_r\}$. Positive floors stabilize concentrated score and radius distributions.

Neighborhoods. Support distances use fixed coordinate-wise medians and IQRs from the initial region. Each IQR is floored at $\max(10^{-8}, 0.1m)$, where m is the median positive initial IQR, or one if none is positive. This scaling affects only neighbor search, which uses the current block and excludes the query itself.

Training split. Select $\lfloor f_c n_t \rfloor$ candidates and split off $\lfloor f_{\text{fit}} |\mathcal{C}_t| \rfloor$ training samples. The split stratifies by confidence approximately equal to or below one, with random tie-breaking and unit weights. Blocks contain at least four samples and both subsets are nonempty. Probes enter ranking and neighborhood statistics, but not gradients.

Replay. After consolidation, admit up to $\max\{1, \lfloor f_r n_t \rfloor\}$ candidates with the smallest support radii, breaking ties by original index; no online confidence cutoff applies. Each step samples up to 256 current and 64 replay entries uniformly without replacement within each set. Samples may recur across steps.

The coefficient β_t controls loss weight, not batch size; an empty buffer sets $\beta_t = 0$. Consolidation counts each training sample once, excluding probes, replay, and repeated steps.

Training contamination. Selection changes which observations contribute to the current-sample DSM term in Equation 15. Let $\widehat{\pi}_t \in (0,1)$ be the block's anomaly fraction, and let a_t^{n} and a_t^{a} denote the fractions of normal and anomalous observations admitted to $\mathcal{T}_t$, respectively, after the train/probe split. For a nonempty training subset, counting its two classes gives

$$\widehat{\pi}_t^{\mathrm{train}} = \frac{\widehat{\pi}_t a_t^{\mathrm{a}}}{(1-\widehat{\pi}_t)a_t^{\mathrm{n}} + \widehat{\pi}_t a_t^{\mathrm{a}}}. \tag{32}$$

Thus $\widehat{\pi}_t^{\mathrm{train}} < \widehat{\pi}_t$ exactly when $a_t^{\mathrm{a}} < a_t^{\mathrm{n}}$. This identity describes the composition of the selected data; labels are used only to define these quantities, not by the method. With uniform replay weights, the data terms in Equation 16 have combined anomaly mass $(1-\beta_t)\widehat{\pi}_t^{\mathrm{train}} + \beta_t \widehat{\pi}_{t-1}^{\mathrm{replay}}$.

B.3 Initialization

The first 99 columns initialize the backbone with AdamW (Loshchilov & Hutter, 2019); Fourier frequencies remain fixed. Training reuses a capped subset sampled without replacement, while all initial vectors contribute to reference statistics and N_0. The initial score quantile α_{init} is 0.9 for Synthetic and 0.8 for SNP and Beach.

Initial confidence q_0 uses the score branch of Eq. (10) with quantile α_{init}. Set $N_0 = \sum_{x \in \mathcal{D}_0} q_0(x)$; capping begins at the first consolidation. Initial replay ranks samples with $q_0 \geq 0.9$, uses stable ties, selects at most 10%, and FIFO-trims to 4096 unit-weight entries. Online admission instead uses support radii.

B.4 Online optimization

Each block resets $A_{t,\ell}$ with Kaiming-uniform initialization and $B_{t,\ell} = 0$, and fits the adapter with a fresh AdamW optimizer. Weight decay and the factor penalty in Eq. (16) both apply.

Probe loss uses two fixed local perturbations per sample and is checked before fitting and every five steps. The best adapter is retained, including the zero-update state. A plateau of three checks with less than 0.5% relative improvement, or loss above 1.2 times the best value, triggers rollback of the adapter and optimizer and an LR reduction by 0.3. Plateau rollback starts after 20 steps. At most two reductions are allowed, with minimum LR 10^{-4}; subsequent plateau or divergence stops fitting after the minimum budget, otherwise fitting continues to 150 steps.

HyperAD experiments use PyTorch 2.2.1 and CUDA 12.1 in FP32 on an RTX 4060 Ti, with four CPU threads.

C Experimental Details

C.1 Data and evaluation

Methods share the block sequence and post-initialization evaluation samples. ESF scores with the restored adapter at scale γ_t, consolidates at ρ_t, then admits replay samples. Updates use initial, current, and retained past data.

HyperAD uses the first 99 columns for initialization and then one column per block. ESF divides every spectrum by the maximum of those initial columns, freezes the divisor, and does not clip later values above one. Baselines retain native values. For a general $H \times W \times C$ cube, scan columns if $H < W$ and rows otherwise.

C.2 Metrics

Ranking metrics. Anomalies are the positive class. AUROC and AUPRC evaluate all post-initialization scores within each stream. For threshold-axis metrics, each method's output scores are min–max scaled to $[0,1]$; constant scores map to zero. This evaluation-only transform preserves rankings and does not enter adaptation or F1 thresholding.

With detection and false-alarm probabilities $P_D(u)$ and $P_F(u)$ at normalized threshold u (Chang, 2021),

$$\begin{aligned} \text{AUROC} &= \text{AUC}(P_D, P_F), \\ \text{AUC-TD} &= \text{AUROC} + \text{AUC}(P_D, u), \\ \text{AUC-BS} &= \text{AUROC} - \text{AUC}(P_F, u), \end{aligned} \tag{33}$$

The threshold-axis areas use trapezoidal integration over finite score thresholds, excluding the artificial threshold above the maximum. AUC-TD and AUC-BS summarize target detectability and background suppression, respectively; both depend on score scaling.

AUPRC integrates precision against recall by the trapezoidal rule, including the endpoint $(R, P) = (0, 1)$. AP instead uses non-interpolated recall increments (scikit-learn developers, n.d.). For precision P_j and recall R_j at distinct thresholds ordered from highest to lowest,

$$\text{AP} = \sum_{j=1}^{J} (R_j - R_{j-1}) P_j, \qquad R_0 = 0, \tag{34}$$

where J is the number of distinct thresholds.

F1 thresholds. For sample-level true positives TP, false positives FP, and false negatives FN,

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

Causal F1 thresholds each block at the $q = 0.99$ quantile of native scores from the ten most recent completed blocks. At the start of a stream, the latest initial-region scores fill the missing history; if fewer than ten blocks are available in total, all initial scores are used. The current block, future scores, and labels do not enter this threshold. Scores strictly above the threshold are anomalous. F1 instead selects the single fixed threshold that maximizes F1 over the evaluated scene or fault stream using its labels; ties favor fewer alarms. It is an offline reference, not a threshold used by the detector.

Aggregation. HyperAD metrics pool post-initialization pixels within each scene. TEP metrics are computed per 1,488-point stream and averaged over 28 faults within each run. Tables 1 and 2 report means across seeds 1–3; Tables 5–7 and 8 also give sample standard deviations. The ablations, detection maps, and per-fault table use seed 1. Both F1 metrics pool decisions within each stream, not across blocks. Anomaly-free blocks remain in evaluation; single-class blocks are excluded only from mean block AUROC.

C.3 HyperAD initial scores

The causal threshold initially uses each detector's own initial scores. For baselines, freeze the detector after 99 native warm-up calls and score the initial columns without updates, retaining native line standardization. RT-CK-RXD uses its boundary mean and inverse covariance; LBL-AD uses the candidate PCA that scored the final warm-up column. Streaming updates remain native.

C.4 TEP configuration

C.4.1 Data and preprocessing

The extended TEP data (Reinartz et al., 2021) provide 85-dimensional observations along a normal Mode-1-to-Mode-2 trajectory. All 28 fault streams share this trajectory and initialization; each replaces 74 observations with independent, mode-matched fault samples. The transition and settling interval remain normal. This tests sparse contamination rather than continuous fault episodes. Initialization uses 512 observations, followed by 23 blocks of 64 and a final block of 16. All methods receive the same 85-coordinate whitening fitted to these 512 samples and the same fixed scale $c = 0.07235$. The target scale was fixed during earlier HyperAD Synthetic development, using whole-cube normalization there; it is not estimated from later TEP observations. No method receives an additional external Z-score transform, and future observations are neither clipped nor used to refit the transform.

ESF settings. The architecture, perturbation schedule, and online fitting settings follow Table 4. Initialization uses 1,250 AdamW updates with batches of 256 sampled uniformly with replacement from all 512 initial observations, at learning rate 3×10^{-4}. The initial confidence quantile is $\alpha_{\text{init}} = 0.9$. For TEP, $N_{\max} = 64$, including the initial cap $N_0 = \min\{64, \sum_{x \in \mathcal{D}_0} q_0(x)\}$; subsequent consolidation uses the same cap.

C.4.2 Baselines and calibration

ARCUS retains the authors' pool updates, batch size 32, LR 10^{-4}, initial/online epochs 5/1, and reliability/similarity thresholds 0.95/0.8. The main comparison uses seeds 1–3.

Final scoring follows each block's update. Window-kNN excludes the query itself; xStream refreshes at native window boundaries. ARCUS adds current-block re-detection to its original pre-update readout. Each baseline is initialized on the shared transformed initial segment and obtains its initial-score reference from eight groups of 64 without updates, retaining native reliability weighting and score standardization. ESF uses its saved initial scores.

C.5 HyperAD ablations

The comparison in Table 3 pairs HyperAD streams and initial states across variants. AUROC, AUPRC, causal F1, F1, AUC-TD, and AUC-BS follow Appendix C.2, with equal-weight averages over Synthetic, SNP, and Beach. Scores from different scenes are not pooled.

All variants start from the same saved initialization within each scene. The four interventions are:

- **No online update.** Freeze the initial score field and score the complete stream without adaptation, consolidation, or replay-bank updates.
- **No confidence selection.** Uniformly sample the same number of candidates from each block and retain the 80/20 fitting/probe split. Adaptive training retains the 20–150-step budget. Confidence-based control and replay admission remain unchanged.
- **Without q^{support}.** Set $q_t^{\text{mix}} = q_t^{\text{score}}$ to form the candidate pool for training, probes, and replay. Candidate counts and h_t-based controls remain unchanged. Replay still selects the smallest support radii within this score-only pool.
- **Without replay.** Set $\beta_t = 0$ while leaving replay-bank maintenance unchanged.

D ADDITIONAL RESULTS

D.1 REPEATED-RUN RESULTS

The scenes and fault streams are fixed across runs; only algorithm randomness changes. The means below are those reported in Tables 1 and 2.

Table 5: **Synthetic (HyperAD).** Three-run mean ± sample standard deviation.

Method	AUROC	AUPRC	Causal F1	F1	AUC-TD	AUC-BS
RX	0.6442±0.0000	0.0507±0.0000	0.0682±0.0000	0.0961±0.0000	0.8255±0.0000	0.5000±0.0000
RT-CK-RXD	0.5816±0.0000	0.0358±0.0000	0.0549±0.0000	0.0684±0.0000	0.7569±0.0000	0.4331±0.0000
RX-BIL	0.5665±0.0013	0.0395±0.0001	0.0573±0.0004	0.0704±0.0010	0.7467±0.0017	0.4114±0.0013
LBL-AD	0.5725±0.0044	0.0151±0.0002	0.0028±0.0007	0.0494±0.0005	1.1610±0.0052	-0.0148±0.0050
ERX	0.7776±0.0302	0.0348±0.0006	0.0443±0.0018	0.0710±0.0112	1.0334±0.0568	0.5867±0.0172
ESF	0.8805±0.0101	0.4589±0.0037	0.1398±0.0008	0.4858±0.0039	1.2760±0.0098	0.7195±0.0140

Table 6: **SNP (HyperAD).** Three-run mean ± sample standard deviation.

Method	AUROC	AUPRC	Causal F1	F1	AUC-TD	AUC-BS
RX	0.9886±0.0000	0.2274±0.0000	0.1254±0.0000	0.3167±0.0000	1.2307±0.0000	0.9104±0.0000
RT-CK-RXD	0.9918±0.0000	0.2840±0.0000	0.1269±0.0000	0.3979±0.0000	1.2754±0.0000	0.9036±0.0000
RX-BIL	0.9917±0.0000	0.2723±0.0012	0.1259±0.0005	0.3861±0.0009	1.2788±0.0002	0.8949±0.0001
LBL-AD	0.9705±0.0001	0.5330±0.0001	0.1299±0.0009	0.6040±0.0014	1.4607±0.0090	0.8261±0.0026
ERX	0.9729±0.0269	0.2433±0.0326	0.1210±0.0014	0.3491±0.0317	1.2607±0.0242	0.8736±0.0353
ESF	0.9933±0.0004	0.8749±0.0017	0.1310±0.0012	0.8282±0.0009	1.9223±0.0023	0.6905±0.0035

Table 7: **Beach (HyperAD).** Three-run mean ± sample standard deviation.

Method	AUROC	AUPRC	Causal F1	F1	AUC-TD	AUC-BS
RX	0.8288±0.0000	0.1727±0.0000	0.0611±0.0000	0.3438±0.0000	1.0278±0.0000	0.7138±0.0000
RT-CK-RXD	0.8057±0.0000	0.2187±0.0000	0.0669±0.0000	0.4196±0.0000	1.0202±0.0000	0.6821±0.0000
RX-BIL	0.8031±0.0002	0.2183±0.0004	0.0665±0.0005	0.4191±0.0009	1.0128±0.0008	0.6834±0.0007
LBL-AD	0.6970±0.0003	0.0768±0.0000	0.0700±0.0003	0.2080±0.0002	0.8332±0.0006	0.5988±0.0002
ERX	0.9387±0.0122	0.1739±0.0099	0.0443±0.0026	0.2963±0.0072	1.1303±0.0177	0.8334±0.0069
ESF	0.9527±0.0010	0.4111±0.0528	0.1209±0.0130	0.5314±0.0069	1.2927±0.0046	0.8401±0.0015

Table 8: **TEP.** Three-run mean ± sample standard deviation of the 28-fault average.

Method	AUROC	AUPRC	Causal F1	F1	AUC-TD	AUC-BS
Window-kNN	0.8968±0.0000	0.6430±0.0000	0.1916±0.0000	0.6402±0.0000	1.0867±0.0000	0.8649±0.0000
Online-iForest	0.5943±0.0085	0.0992±0.0092	0.1387±0.0122	0.1612±0.0031	0.8079±0.0280	0.4203±0.0094
ERX	0.5589±0.0053	0.1154±0.0166	0.0609±0.0043	0.1425±0.0156	1.1297±0.0042	0.0395±0.0139
xStream	0.6674±0.0020	0.0856±0.0088	0.0841±0.0051	0.1955±0.0018	1.3312±0.0057	0.1929±0.0032
ARCUS-RSRAE	0.5621±0.0487	0.3567±0.0426	0.2599±0.0402	0.4124±0.0444	1.0508±0.0856	0.1346±0.0581
ARCUS-RaPP	0.9006±0.0242	0.4757±0.1007	0.2269±0.0564	0.5778±0.0821	1.6257±0.0461	0.3792±0.0351
ESF	0.9496±0.0020	0.7788±0.0051	0.3223±0.0018	0.8054±0.0025	1.7308±0.0059	0.7940±0.0074

D.2 CONTROLLED SYNTHETIC DIAGNOSTICS

We use a 16-dimensional stream with two-dimensional latent normal structure to examine how the field changes as backgrounds enter and leave the scan. The initial set contains 512 clean samples from mode A. Across 24 blocks of 128 samples, A transitions to B in blocks 5–8, B occupies blocks 9–18, and A returns as a 5% minority in blocks 19–24. Paired streams contain either 6 or 26 off-normal anomalies per block (nominally 5% or 20%). Fixed, independent A, B, and anomaly probes are excluded from online training, selection, and replay.

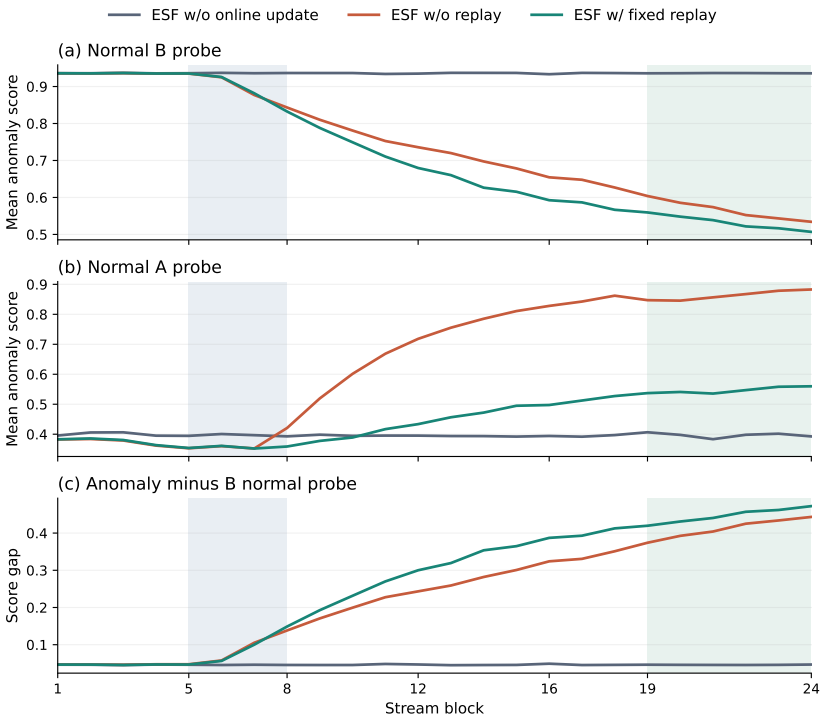

Figure 4: **Evolution of held-out synthetic probes.** Mean raw anomaly scores for normal B and A, and the anomaly/normal-B score gap in the 5% stream. Curves show ESF w/o online update, w/o replay, and w/ fixed replay (weight 0.05). Shading marks the A-to-B transition (blocks 5–8) and A's return (blocks 19–24).

Stream construction. Standard Gaussian latent coordinates z are embedded in two-dimensional orthonormal bases in $\mathbb{R}^{16}$. Both bases derive from a common random orientation; B adds Gaussian perturbations of standard deviation 0.35 before orthonormalization. Their offsets differ by 2.0 along an axis orthogonal to the common base. Component $c \in \{0, 1\}$ adds $0.3\sin((1.5 + 0.15c)z_1 + 0.005t)$ along a unit direction orthogonal to its base. A shared translation of $0.25\sin(0.05t)$ affects the first four coordinates; isotropic Gaussian observation noise has standard deviation 0.03. Anomalies add a component-specific offset of length 4.0 orthogonal to the corresponding base plane. Both contamination levels share latent samples and geometry, with nested anomaly positions. Initialization uses 10 epochs, $N_0 = 512$, and the 51 initial samples with smallest support radii for replay. Each held-out probe group contains 64 samples.

New normal mode. Under the full ESF policy in the 5% stream, the selected 64-sample candidate pool contains 0, 4, 36, and 64 B samples in blocks 5–8: B enters the update as it becomes prevalent. By block 24, the held-out B score is 0.531 under this policy, versus 0.936 without online updates. In Figure 4, the w/o replay and fixed-replay variants reach B scores of 0.534 and 0.507, with anomaly-B/normal-B gaps of 0.443 and 0.472, respectively.

Contamination. Under persistent 20% contamination, 624 anomalies occur in the stream. The full ESF policy admits 3 of them among 1,224 current training samples and none into replay; the equal-size random-selection variant admits 230. Table 9 shows the resulting loss of anomaly separation under random selection. In a separate matched fork of the full policy that changes only block 12 from 6 to 26 anomalies, neither branch admits an anomaly to training or replay, and their held-out B scores differ by 0.010 at block 24.

Table 9: **Persistent contamination in the synthetic stream.** The two policies use equal-size current training sets; AUROC and AUPRC pool the 24 blocks. Probe scores are means on held-out B-origin anomalies during established B.

Current selection	Anomalies trained	Anomaly-B score	AUROC	AUPRC
Confidence-based	3 / 1,224	0.9712	0.9884	0.9751
Random	230 / 1,224	0.8135	0.7057	0.4977

Returning normal mode and replay. At block 24, the held-out A score is 0.883 for w/o replay and 0.560 for fixed replay (Figure 4). Separately, under the full ESF policy, all 34 returning-A normals in blocks 19–24 rank below the 36 actual anomalies in deployed scores (conditional AUROC and AUPRC both 1.000). Figure 5 forks the full policy at block 21 to isolate the immediate effect of replay weight. Raising it from 0 to 0.4 lowers the mean score of six returning-A samples from 0.936 to 0.671 while six anomaly scores remain near 0.977. The score difference largely dissipates after the original adaptive weighting resumes in blocks 22–24.

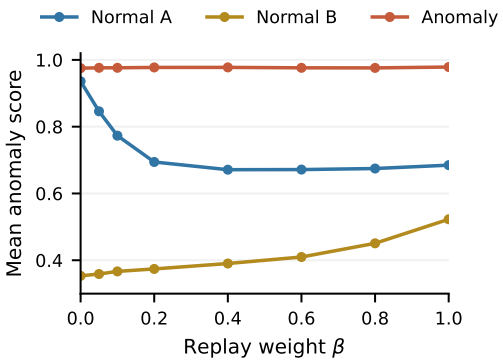

Figure 5: **Replay weight in block 21.** Deployed scores for returning A, current B, and B-origin anomalies from a shared checkpoint.

D.3 TEP RESULTS

Table 10 reports seed-1 ESF results for each fault. In that run, it leads causal F1 on 19 faults and F1 on 14, including ties. Faults 15, 23, 28, and 16 have its lowest AUPRC values, between 0.1974 and 0.2997.

Table 10: **Per-fault TEP results for ESF (seed 1).** Causal F1 uses the past-only $q = 0.99$ threshold; F1 uses the best fixed threshold.

Fault	AUROC	AUPRC	Causal F1	F1	Fault	AUROC	AUPRC	Causal F1	F1
01	1.0000	1.0000	0.3115	1.0000	15	0.7509	0.1974	0.0702	0.3171
02	0.9985	0.9720	0.3731	0.9308	16	0.7457	0.2997	0.0877	0.4110
03	1.0000	1.0000	0.4648	1.0000	17	0.9928	0.9093	0.4058	0.8281
04	1.0000	1.0000	0.4460	1.0000	18	0.9379	0.6044	0.2400	0.7143
05	1.0000	1.0000	0.4539	1.0000	19	0.9930	0.8817	0.2880	0.8313
06	1.0000	1.0000	0.7471	1.0000	20	0.9920	0.8300	0.3256	0.8343
07	1.0000	1.0000	0.2698	1.0000	21	0.9982	0.9860	0.3704	0.9726
08	0.9927	0.8289	0.4058	0.8555	22	0.9985	0.9768	0.4429	0.9429
09	0.9988	0.9818	0.4058	0.9504	23	0.7779	0.2690	0.0696	0.3796
10	0.9990	0.9849	0.4203	0.9510	24	0.8276	0.3146	0.1667	0.4143
11	0.9990	0.9837	0.4397	0.9577	25	0.9717	0.6423	0.2540	0.7303
12	0.9980	0.9775	0.3609	0.9504	26	0.9654	0.6211	0.2560	0.7273
13	0.9996	0.9933	0.4478	0.9548	27	0.9402	0.4367	0.1833	0.6339
14	0.9921	0.7788	0.2623	0.8706	28	0.7728	0.2901	0.1034	0.3797

D.4 HyperAD detection maps

Figures 6–8 show seed-1 scores from all six methods and HyperAD scene previews (Garske et al., 2025).

- **Display:** full-scene orientation and aspect ratio; independent linear min–max scaling over post-initialization scores, without clipping or smoothing.
- **Markers:** white GT pixels denote anomalies; dashed lines mark the initial 99 columns. Available initial values are retained; missing baseline scores appear black.

Scans proceed left to right. Brightness is relative within each map, not a shared decision threshold.

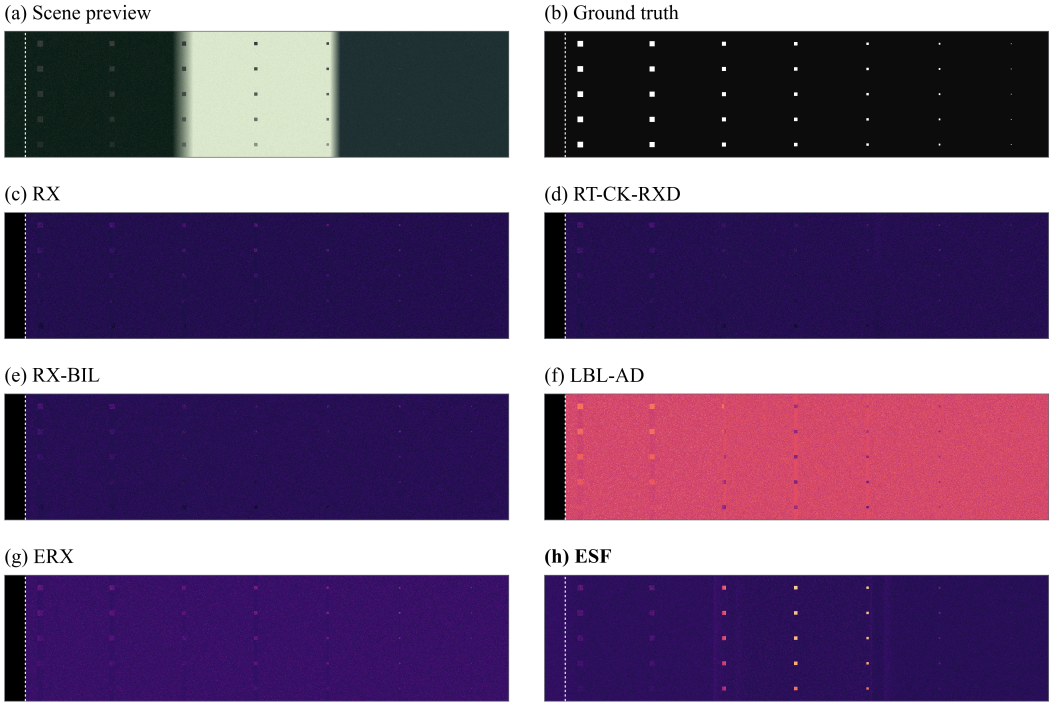

Figure 6: **Synthetic.** Scene preview, ground truth, and six detection maps across successive backgrounds.

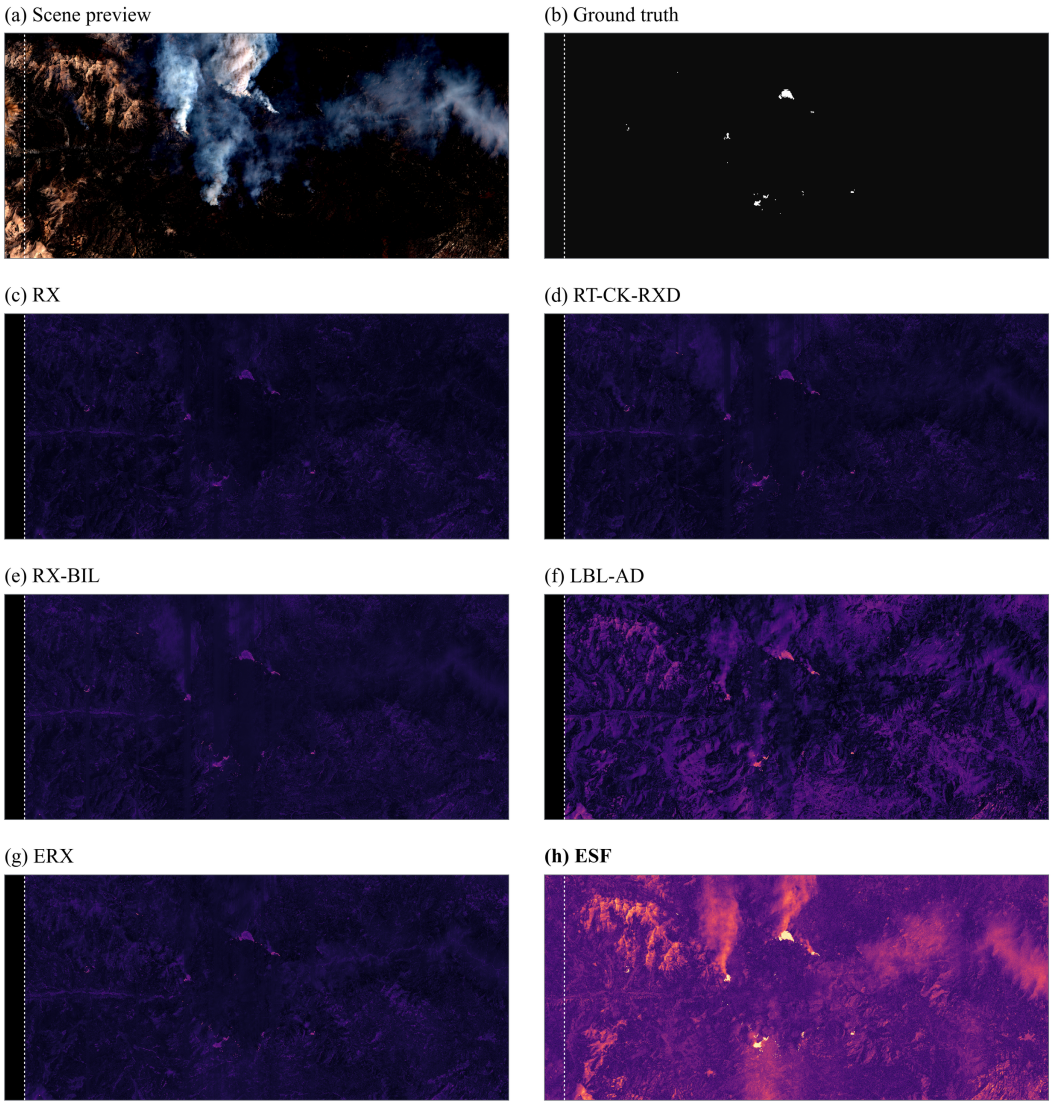

Figure 7: **SNP.** False-color preview, ground truth, and detection maps showing fire-region and terrain responses.



Figure 8: **Beach.** Scene preview, ground truth, and detection maps across vegetation, sand, and shoreline.

E REMOTE-SENSING APPLICATIONS

We present qualitative vessel and fire-point examples using independent models with the HyperAD settings. Initialization uses 10 epochs and at most 50,000 spectra from the first 99 columns for vessels or 99 rows for fire points; subsequent columns or rows form individual blocks. These retrospective visualizations retain their original full-scene scaling and signed fire spectra. No target labels enter adaptation.

Figures 9 and 10 pair RGB views with ESF maps, displayed left to right in scan order. Each map uses its complete valid score range, including initialization, without smoothing or local contrast stretching. Brighter colors indicate higher scores.

E.1 MARITIME VESSEL DETECTION

The Wyvern strips[1] cover Suez, Chesapeake Bay, Bahrain, and Dongshan Bay. Regions are selected from RGB views before detection. Rotation uses native spectra without interpolation, processing repeated pixels only at their first occurrence.

ESF produces localized vessel responses across varying water backgrounds. Suez shows clear vessel groups; some small vessels in Chesapeake Bay and Dongshan Bay respond weakly. Bahrain retains additional responses along the water-color transition.

E.2 FOREST FIRE-POINT DETECTION

The four complete AVIRIS-C flightlines from Fire-D v5[2] show localized responses near smoke-bearing regions amid terrain and agricultural variation. Scene `20190821_p00r27` has a compact response cluster; `20190806_p00r18` also responds away from the main smoke region. Visible stripes in `20190807_p00r14` produce additional responses. The maps indicate spectral anomalies, not complete fire boundaries.

[1] `https://opendata.wyvern.space/`

[2] `https://zenodo.org/records/15428368`

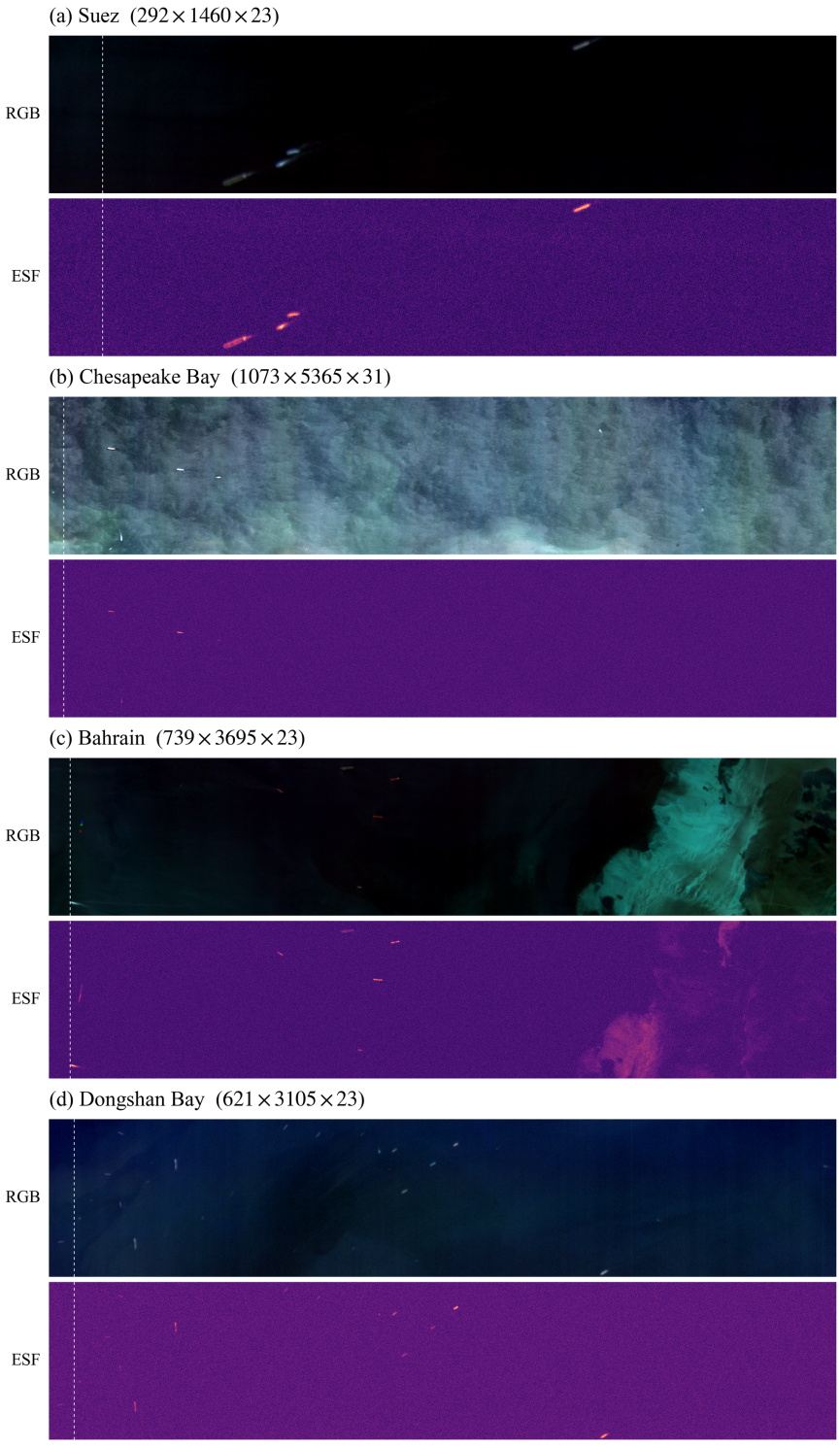

Figure 9: **Maritime vessel examples.** RGB above ESF; dimensions give the region's $H \times W \times C$. Dashed lines mark initialization. RGB: © 2025 Wyvern Incorporated for (b); © 2024 for (a), (c), and (d). All Rights Reserved.

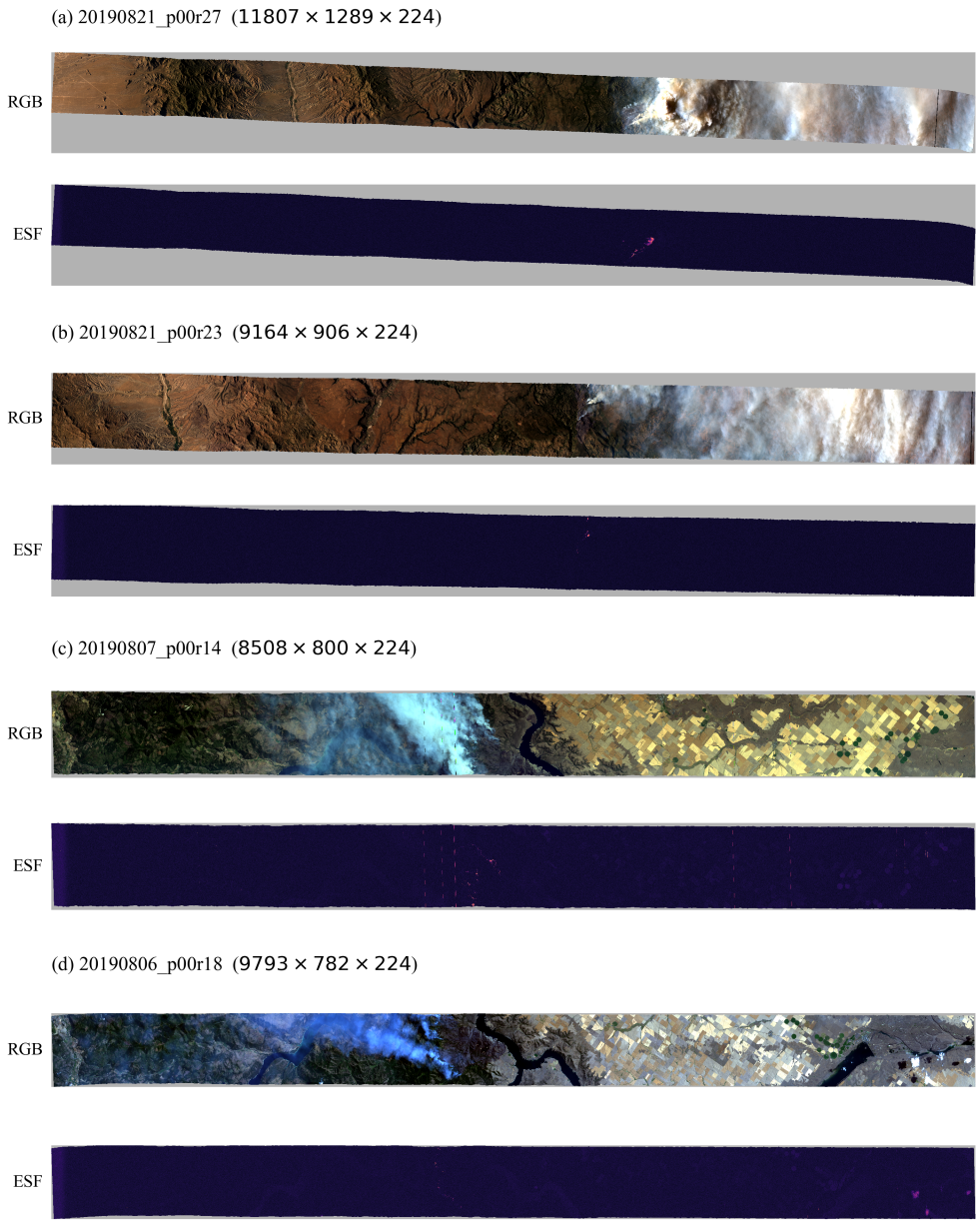

Figure 10: **Forest fire-point examples.** RGB above ESF; dimensions give the original $H \times W \times C$. Complete flightlines are rotated for display. Gray denotes invalid pixels; stripe artifacts in (c) are retained.

F LIMITATIONS

Sustained high contamination. ESF learns from unlabeled data through self-supervised score matching, relying on normal observations to dominate the incoming blocks. Persistent, high-proportion anomalies can acquire strong neighborhood support and enter adaptation, causing the field to absorb them as background rather than suppress them.

Dependence on manifold structure. ESF's anomaly score exploits the local geometry of backgrounds concentrated near low-dimensional manifolds. When high-dimensional backgrounds lack this structure or exhibit it only weakly, score-direction consistency can lose its ability to distinguish anomalies from background.

Computational cost. Our current implementation has higher per-block runtime than statistical detectors such as ERX (Garske et al., 2025): repeated score evaluations and online adapter optimization dominate its cost. Reducing this overhead while retaining adaptation quality is an important direction for future work.