COARSE TEMPORAL FEEDBACK FOR TEST-TIME ADAPTATION IN STREAMING TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) can help time-series anomaly detectors cope with evolving normal behavior during deployment. Yet label-free TTA must decide which test observations are safe for adaptation using model-derived signals that may themselves become unreliable under distribution shift. External feedback offers an additional source of adaptation evidence, but in deployment it may be available only periodically and at a coarse temporal granularity. We therefore study a setting in which each completed interval is summarized by one binary response that can influence only future predictions: a negative response certifies the interval as normal, whereas a positive response reveals only that an anomaly occurred somewhere within it. We introduce Feedback-Interval Test-Time Adaptation (FITTA), a minimal adaptation method for this strictly chronological setting. FITTA reprocesses intervals confirmed to be normal for adaptation, penalizes reductions in the interval maximum anomaly score relative to the frozen source detector after positive feedback, and uses the latest response to set the resolution of the next feedback interval. With interval feedback simulated from benchmark labels and the seven benchmark families weighted equally, FITTA raises AUPRC from 0.327 to 0.551 and VUS-PR from 0.248 to 0.506 relative to the frozen detector. These results show that coarse existential feedback can provide substantial adaptation utility for subsequent detection without anomaly localization. A matched factorial analysis further shows that most of this utility comes from adaptation on intervals certified as normal, while positive-evidence preservation and adaptive resolution provide a smaller, non-additive refinement rather than independent additive gains.

1 INTRODUCTION

Time-series anomaly detection (TSAD) is widely used in continuous monitoring systems, where detectors commonly learn a notion of normality from historical observations and identify subsequent deviations (Cook et al., 2020; Wen et al., 2022; Blázquez-García et al., 2022). In long-running deployments, however, normal operating conditions can evolve as workloads change, sensors drift, or previously unseen benign modes emerge. Such shifts can cause a fixed detector to assign high anomaly scores to legitimate observations, producing excessive false positives (Kim et al., 2024; 2026). Periodic recalibration or retraining can accommodate such changes, but may lag behind an evolving stream because they require a separate cycle of model updating and redeployment. This motivates test-time adaptation (TTA), which updates the deployed detector during inference as new data arrive (Wang et al., 2021; 2025b; Kim et al., 2024; 2026).

Once a detector adapts on the test stream, incoming observations are no longer only data to be scored, but may also serve as evidence for subsequent model updates. The central difficulty is therefore whether such adaptation evidence can be trusted. Label-free TTA must infer this evidence from the same shifted target stream that the detector is trying to adapt to. This creates a self-referential risk: under distribution shift, its own scores or representations may be unreliable guides for deciding which observations are safe to use for adaptation. A mistaken decision is then more consequential than an isolated prediction error, because it changes the model state used for subsequent observations and can cause true anomalies to be absorbed as normal or otherwise degrade anomaly discrimination.

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

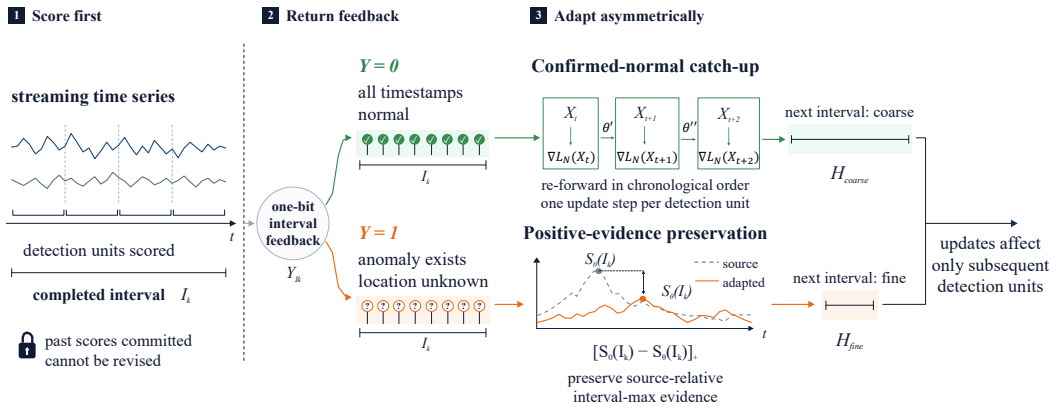


Figure 1: Overview of the coarse-feedback setting and FITTA. Each interval is fully scored before one existential feedback bit is returned, and committed scores are never revised. Negative feedback certifies the interval as normal and enables chronological catch-up with one update per detection unit; positive feedback reveals anomaly existence without localization and triggers source-relative interval-max preservation. The returned bit determines the next feedback resolution, while all updates affect only subsequent detection units.

This motivates seeking adaptation evidence that does not rely solely on the detector's own judgments under shift.

External feedback provides such an additional source of adaptation evidence. Prior TSAD methods have used external labels for selected samples or endpoints, including in streaming settings (Wang et al., 2022; Zhou et al., 2024). Even when feedback is sparse, these labels are attached to specific queried observations or endpoints. In practice, however, obtaining such feedback can require substantial human labeling effort, and feedback may be collected only periodically rather than continuously (Liu et al., 2015; Bodor et al., 2022). Existing approaches address this constraint mainly by reducing how many observations are queried, while each returned label still refers to a specific sample or endpoint. This leaves the temporal specificity of each feedback response itself largely unquestioned.

This raises a separate question: how much temporal localization must each feedback response provide to support adaptation? We study a deliberately coarse setting that removes point-level localization altogether. After a contiguous interval has been scored, the feedback provider returns one binary response: a negative response certifies every timestamp in the interval as normal, whereas a positive response reveals only that at least one anomaly occurred somewhere within it, without identifying where or how many. The response is returned only after the interval has ended, cannot revise previously committed scores, and can influence only subsequent predictions. This setting creates two design questions. First, how should adaptation use the asymmetric information carried by the two responses? Second, how should the feedback interval be chosen to balance interaction frequency against the temporal ambiguity of positive feedback?

To study these questions, we introduce Feedback-Interval Test-Time Adaptation (FITTA), a minimal adaptation method for this setting (Figure 1). Following a negative response, FITTA performs adaptation on the interval certified as entirely normal. Following a positive response, FITTA disables normal learning and penalizes reductions in the interval maximum anomaly score relative to the frozen source detector. The maximizing timestamp serves only as a high-score witness and is not interpreted as the anomaly location. The most recent response then determines the resolution of the next interval, using finer feedback after a positive response and coarser feedback after a negative response.

Across seven benchmark families, FITTA increases AUPRC from 0.327 to 0.551 and VUS-PR from 0.248 to 0.506 when the families are weighted equally. These results show that coarse existential feedback can provide substantial adaptation utility for subsequent detection despite revealing no

anomaly locations. A matched factorial analysis further shows that most of this utility comes from adaptation on intervals certified as normal by negative feedback, while positive-evidence preservation and adaptive temporal resolution provide a smaller, interacting refinement. A matched point-localization oracle analysis yields only modest additional aggregate utility under the tested update objectives, while also revealing non-additive interactions between two uses of exact localization. At matched response counts, FITTA also outperforms the deterministic control and all 20 pre-specified random open-loop schedules in aggregate, indicating that response count alone does not explain the adaptive result. The main utility gain transfers to USAD, a second reconstruction-based detector.

Our contributions are threefold:

- **Setting.** We formulate test-time adaptation for strictly chronological TSAD under coarse existential interval feedback, where each returned bit summarizes a completed interval without revealing anomaly locations and can affect only future predictions. This setting explicitly separates feedback frequency from temporal localization and makes interval resolution part of the adaptation problem.
- **Method.** We introduce FITTA, a minimal adaptation method tailored to the asymmetric semantics of interval feedback. Negative responses enable confirmed-normal adaptation, positive responses trigger source-relative preservation of interval-level anomaly evidence, and the latest response determines the temporal resolution of the next feedback interval.
- **Characterization.** Across seven benchmark families, we show that coarse existential feedback provides substantial adaptation utility for subsequent detection. Matched factorial analysis reveals that most of this utility comes from intervals certified as normal by negative feedback, while positive-evidence preservation and adaptive resolution provide a smaller interacting refinement. Point-localization oracle diagnostics characterize the remaining value of exact temporal localization under the tested update objectives, while matched-response controls show that response count alone does not explain the adaptive result. The main utility gain also transfers to a second reconstruction-based detector.

2 Related Work

To position our setting, we distinguish prior work along three dimensions: the source of adaptation evidence, the temporal specificity of external feedback, and the role of coarse temporal labels in model updating.

Test-Time Adaptation for Time-Series Anomaly Detection. Test-time adaptation has been introduced to TSAD to track evolving normal behavior under distribution shift. Recent TSAD-TTA methods such as M2N2 (Kim et al., 2024) and CANDI (Kim et al., 2026) adapt without external test-time labels, but differ in how they identify adaptation evidence. M2N2 updates the detector using test observations predicted to be normal, whereas CANDI curates candidate adaptation samples using anomaly scores and latent-space similarity to source-side references and restricts updates to a lightweight module. Thus, both methods derive adaptation evidence from unlabeled test observations and model- or source-derived signals. In contrast, our setting introduces externally returned coarse interval feedback after scoring to guide model updates that affect only subsequent observations.

Feedback-Driven and Active Time-Series Anomaly Detection. External feedback has also been used to improve TSAD through active and interactive learning. ATAD (Zhang et al., 2019), SLA-VAE (Huang et al., 2022), and Active-MTSAD (Wang et al., 2022) acquire labels for selected target instances, while operator-feedback methods refine detectors from verified alarms or point-level corrections (Bodor et al., 2022; Guo et al., 2022). BALAD (Zhou et al., 2024) extends active querying and model updates to time-series data streams, and Han et al. (2026) use queried labels to fine-tune reconstruction-based anomaly detectors. Thus, external feedback for TSAD is not itself new. However, the returned feedback in these settings is primarily attached to specific queried samples, endpoints, alarms, or point-level corrections. Our setting instead removes this point-level temporal localization and returns one response for a completed interval.

Weak Temporal Supervision. Coarse temporal labels with existential semantics are likewise established in weakly supervised anomaly detection. WETAS (Lee et al., 2021) learns temporal localization from sequence-level anomaly labels, NRdetector (Wang et al., 2025a) uses weak segment labels whose positive and negative semantics closely match ours, and TreeMIL (Liu et al., 2024) studies point-level TSAD under inexact temporal supervision. These works show that anomaly supervision need not identify exact anomalous timestamps. Their coarse labels, however, are used as offline supervision for learning or localization, whereas our interval response arrives only after the interval has been scored and can influence only future adaptation. The distinction is therefore not the existential interval-label semantics themselves, but their chronological role in test-time adaptation.

3 Problem Formulation

We consider a multivariate time-series stream $\mathbf{x}_1, \mathbf{x}_2, \ldots$, where $\mathbf{x}_r \in \mathbb{R}^C$ denotes the observation at raw timestamp r. Let f_{θ_0} be a reconstruction-based detector pre-trained on source-domain normal data. During deployment, we use θ to denote the current mutable model state, initialized from θ_0.

Because the detector processes temporal windows or blocks rather than isolated timestamps, we treat each processed window or block as a detection unit X_t and denote the raw timestamps covered by that unit by $\mathcal{R}(X_t)$. Detection units are processed in chronological order. When X_t is processed, the current model produces timestamp-level anomaly scores $\{s_\theta(r) : r \in \mathcal{R}(X_t)\}$. Once produced, these scores are committed and are never revised.

We partition the stream of detection units online into successive non-overlapping feedback intervals $I_1, I_2, \ldots$. Interval I_k contains H_k consecutive detection units, with $\mathcal{R}(I_k) = \bigcup_{X_t \in I_k} \mathcal{R}(X_t)$. All detection units in I_k are scored before any feedback for that interval is returned, and the deployment state is held fixed while the interval is being scored. Only after I_k has been completed does the feedback provider return one binary response:

$$Y_{I_k} = \begin{cases} 0, & y_r = 0 \text{ for all } r \in \mathcal{R}(I_k), \\ 1, & y_r = 1 \text{ for at least one } r \in \mathcal{R}(I_k), \end{cases}$$

where $y_r \in \{0, 1\}$ is the latent point-level anomaly state used only to define the feedback semantics and is never observed by the detector.

The two responses carry asymmetric information. A negative response, $Y_{I_k} = 0$, certifies that every timestamp in the completed interval is normal. A positive response, $Y_{I_k} = 1$, reveals only that at least one anomaly occurred somewhere within the interval, without identifying its location or the number of anomalous timestamps. After the response is returned, the model may be updated using the completed interval and its feedback, but any resulting update can affect only subsequent detection units; all previously committed scores remain unchanged. We assume that feedback is available at interval completion before the next interval is scored and do not model additional response or processing delay.

Under this protocol, the objective is to use coarse interval feedback to improve detection on the future stream while limiting feedback interactions. The interval length H_k determines the temporal resolution of each response: shorter intervals provide more localized feedback but require more frequent responses, whereas longer intervals reduce interaction frequency while making positive feedback more temporally ambiguous.

4 FITTA: Feedback-Interval Test-Time Adaptation

FITTA is a minimal adaptation method for the setting defined in Section 3. After each completed feedback interval, it uses the returned bit in two ways: to determine how the detector is adapted and to set the temporal resolution of the next feedback interval.

4.1 Feedback-Conditioned Adaptation

FITTA operates on the timestamp-level anomaly score $s_\theta(r)$ defined in Section 3. For the primary autoencoder, this score is the channel-averaged squared reconstruction error $s_\theta(r) =$

$\frac{1}{C}\sum_{c=1}^{C}(x_{r,c}-\hat{x}_{\theta,r,c})^2$. In addition to the mutable deployment model, FITTA retains the source-trained detector f_{θ_0} as an immutable reference.

Confirmed-normal catch-up. A negative response $Y_{I_k}=0$ certifies every timestamp in the completed interval as normal. FITTA therefore reprocesses its detection units in chronological order, taking one optimization step per unit with

$$\mathcal{L}_{\mathrm{N}}(X_t;\theta)=\frac{1}{|\mathcal{R}(X_t)|}\sum_{r\in\mathcal{R}(X_t)}s_\theta(r).$$

This uses the newly certified interval as normal adaptation data.

Positive-evidence preservation. A positive response $Y_{I_k}=1$ establishes only that an anomaly occurred somewhere in the interval, so FITTA disables confirmed-normal adaptation for that interval. Instead, it compares the interval-level maximum score under the deployment and immutable source models:

$$S_\theta(I_k)=\max_{r\in\mathcal{R}(I_k)}s_\theta(r),\qquad S_0(I_k)=\max_{r\in\mathcal{R}(I_k)}s_{\theta_0}(r),$$

and uses the zero-margin preservation loss

$$\mathcal{L}_{\mathrm{P}}(I_k)=\left[S_0(I_k)-S_\theta(I_k)\right]_+.$$

If the loss is positive, FITTA performs one preservation update; otherwise, no update is made. The maximum serves only as an existential high-score witness: its maximizing timestamp is not interpreted as the anomaly location. The preservation signal is only as informative as the source-model interval maximum.

4.2 Adaptive Temporal Resolution

To address the trade-off between feedback resolution and interaction frequency described in Section 3, FITTA switches between two feedback interval lengths, H_{coarse} and H_{fine}, with $H_{\mathrm{fine}}<H_{\mathrm{coarse}}$, and starts with H_{coarse}. After feedback for I_k is returned, FITTA sets $H_{k+1}=H_{\mathrm{fine}}$ if $Y_{I_k}=1$ and $H_{k+1}=H_{\mathrm{coarse}}$ otherwise, thereby using finer feedback after a positive response and coarser, lower-frequency feedback after a negative one. The rule depends only on the most recently returned feedback bit. It assumes that the feedback provider can support both interval lengths. Detailed re-forwarding, optimizer-state, and terminal-interval semantics are given in Appendix A.

5 Experiments

We first evaluate detection utility and transfer to a second reconstruction-based detector. We then use matched analyses to identify the sources of utility, characterize their dependence on temporal resolution, and assess the remaining value of exact point localization. Finally, we examine the interaction–utility trade-off under matched response counts and sensitivity to directional feedback errors.

5.1 Experimental Setup

Data and protocol. Our evaluation comprises GHL, CATSv2, PUMP, SMD, PSM, SWAT, and GECCO. Each entity is evaluated independently over three seeds. Detection uses non-overlapping windows of length $L=12$ and stride 12 with online batch size one. Evaluation is strictly chronological: all windows in a feedback interval are scored before its feedback is returned or any feedback-dependent adaptation occurs, and committed scores are immutable. In FITTA, benchmark annotations determine only the existential interval feedback; point-level labels are used for evaluation and retrospective diagnostics, and are exposed to the updater only in the oracle analyses.

Models and comparisons. The primary detector is a three-layer autoencoder (AE). FITTA uses $(H_{\mathrm{coarse}},H_{\mathrm{fine}})=(20,10)$ and updates the full AE with AdamW at learning rate 10^{-4}. Frozen, M2N2 (Kim et al., 2024), and CANDI (Kim et al., 2026) use harmonized checkpoints, preprocessing, and evaluation support where applicable; M2N2 and CANDI receive no external feedback.

Table 1: Detection utility across the seven evaluated benchmark families. Each panel reports family means and their equal-weight macro average. M2N2 and CANDI use no external feedback; BALAD uses point-specific endpoint feedback under its native protocol. Comparisons are therefore not uniformly matched across methods.

(a) Point AUPRC

Method	GHL	CATSv2	PUMP	SMD	PSM	SWAT	GECCO	Macro
Frozen	0.014	0.124	0.237	0.445	0.549	0.717	0.203	0.327
M2N2	0.011	0.186	0.202	0.413	0.398	0.174	0.635	0.288
CANDI	0.014	0.124	0.237	0.445	0.549	0.717	0.203	0.327
BALAD	**0.090**	0.161	0.214	0.243	0.358	0.208	0.536	0.259
FITTA	0.066	**0.283**	**0.800**	**0.485**	**0.677**	**0.854**	**0.692**	**0.551**

(b) VUS-PR

Method	GHL	CATSv2	PUMP	SMD	PSM	SWAT	GECCO	Macro
Frozen	0.014	0.105	0.249	0.410	0.503	0.418	0.035	0.248
M2N2	0.012	0.173	0.199	0.361	0.393	0.167	0.484	0.256
CANDI	0.014	0.105	0.249	0.410	0.503	0.418	0.035	0.248
BALAD	**0.086**	0.148	0.215	0.214	0.325	0.201	0.437	0.232
FITTA	0.061	**0.269**	**0.799**	**0.454**	**0.650**	**0.739**	**0.572**	**0.506**

BALAD (Zhou et al., 2024) uses point-specific endpoint feedback with its native detector and protocol, so it is contextual rather than a strictly matched baseline. USAD (Audibert et al., 2020) supplies a second reconstruction-based backbone.

Metrics and aggregation. The primary metric is Point AUPRC, with VUS-PR as a secondary metric. Metrics are computed per entity and seed, averaged within each family, and then averaged across the seven families with equal family weight. Interaction frequency is reported as returned responses per 1,000 evaluated raw timestamps using the same aggregation structure.

5.2 Detection Utility and Cross-Backbone Transfer

Table 1 shows that FITTA raises the equal-family macro Point AUPRC from 0.327 to 0.551 and VUS-PR from 0.248 to 0.506 relative to Frozen. FITTA exceeds Frozen at the family-mean level on all seven families for both metrics.

CANDI is evaluated through a harmonized implementation under the common chronological protocol rather than its native experimental pipeline. It performs adaptation updates in 97.2% of evaluated entity–seed deployments, while its full-precision family and macro metrics differ only slightly from Frozen and therefore coincide after three-decimal rounding.

To test whether the main utility gain is specific to the primary AE, we also instantiate FITTA on USAD, a second reconstruction-based detector. FITTA reaches 0.5470 Point AUPRC and 0.4964 VUS-PR with USAD, compared with 0.3255 and 0.2423 for the corresponding Frozen model, and also exceeds the tested fixed-$H = 20$ and matched-budget controls in the equal-family macro. This supports transfer of the main utility gain to a second reconstruction-based detector. Full family-level and control results are reported in Appendix F.

5.3 Sources of Utility and Resolution Dependence

We first use a matched 2×2 design to identify the main source of utility and the interaction between preservation and temporal resolution. The design crosses preservation (off/on) with temporal resolution (fixed $H = 20$/adaptive). Fixed $H = 20$ with confirmed-normal catch-up alone reaches 0.5411 Point AUPRC, versus 0.3267 for Frozen, accounting for 95.5% of the Frozen-to-FITTA AUPRC gain. Thus, adaptation on intervals certified as normal accounts for most of the observed utility in this design. Adding preservation at fixed $H = 20$ contributes only $+0.0032$ AUPRC. Adaptive resolution without preservation is -0.0063 below fixed $H = 20$, whereas preservation contributes $+0.0164$ under adaptive resolution and adaptive resolution contributes $+0.0069$ when preservation

Table 2: Component attribution and preservation dependence on temporal resolution. Panel (a) reports Point AUPRC / VUS-PR for the matched 2×2 factorial; panel (b) reports preservation ON–OFF across fixed resolutions. Deltas are computed from unrounded values.

(a) Factorial attribution			*(b) Preservation effect across fixed H*					
Cadence	Normal-only	+ Pres.		$H = 5$	$H = 10$	$H = 20$	$H = 40$	$H = 80$
Fixed $H = 20$	0.541 / 0.493	0.544 / 0.498	Δ AUPRC	+0.034	+0.017	+0.003	-0.002	-0.003
Adaptive	0.535 / 0.487	**0.551 / 0.506**	Δ VUS-PR	+0.037	+0.020	+0.005	-0.001	-0.002

Table 3: Consistency of the matched Point-AUPRC component contrasts. Family signs count positive/negative family means; seed signs correspond to the three deployment seeds.

Contrast	Macro Δ	Families $+/-$	Seed signs
Fixed normal-only − Frozen	+0.2144	7/0	$+/+/+$
Preservation, fixed	+0.0032	5/2	$-/+/+$
Preservation, adaptive	+0.0164	5/2	$+/+/+$
Adaptive, no preservation	-0.0063	3/4	$-/-/-$
Adaptive, + preservation	+0.0069	6/1	$-/+/+$

is enabled. The resulting descriptive interaction is $+0.0132$ AUPRC, and VUS-PR has the same qualitative ordering.

The family- and seed-level contrasts in Table 3 reinforce this attribution. The confirmed-normal gain is positive across all seven families and all three seeds, whereas the smaller preservation and adaptive-resolution effects are substantially more heterogeneous. Full family-level contrasts are reported in Appendix H.

Panel (b) further shows that preservation is resolution-dependent even under fixed schedules. Its aggregate effect is positive at $H = 5$, 10, and 20, but slightly negative at $H = 40$ and 80 on both metrics. This five-point analysis does not identify the cause of the sign change; complete ON/OFF results and family/seed sign counts are reported in Appendix D.

Retrospective diagnostics provide partial context for this interaction but do not identify its mechanism. Under adaptive resolution, positive-interval anomaly occupancy increases from 0.6380 to 0.7072 and active preservation opportunities increase from 0.3352 to 0.4339 per 1,000 timestamps, while the activation rate conditional on a positive interval decreases from 0.6138 to 0.5551. Confirmed-normal coverage also increases from 0.8582 to 0.8664, so reduced normal coverage is not supported as an explanation for the adaptive normal-only decrement.

Recent feedback is also strongly temporally structured: under adaptive resolution, the next interval is positive with probability 0.739 after a positive response, compared with 0.060 after a negative response. This pattern supports temporal concentration of positive feedback, while retrospective score-shift analyses do not support a universal interpretation that preservation improves detection by preventing future anomaly-score attenuation. Full diagnostics are reported in Appendix G.

5.4 Point-Localization Oracle Diagnostics

We next ask how much additional utility becomes available if positive feedback also reveals exact point-level anomaly locations after the interval has already been scored. We compare FITTA with three matched oracle variants that use this information either to restrict the preservation witness to true anomalous timestamps, to recover normal timestamps inside positive intervals for confirmed-normal catch-up, or to do both (Table 4). All four arms share identical feedback intervals, response counts, and evaluation support.

Restricting the preservation witness to true anomalous timestamps raises Point AUPRC from 0.5511 to 0.5547, while using the revealed normal-point mask for additional confirmed-normal catch-up reaches 0.5603. Combining both uses yields 0.5538 rather than a further gain; the descriptive 2×2 interaction is -0.0100 AUPRC. Thus, under the tested update objectives, exact point localization

Table 4: Point-localization oracle diagnostics under the final FITTA protocol. All arms share identical feedback intervals, response counts, and evaluation support. Exact point labels are revealed only after a positive interval has been fully scored and can affect future updates only.

Positive-interval localization	Point AUPRC	VUS-PR
None (FITTA)	0.5511	0.5063
Anomaly witness	0.5547	0.5111
Normal-point mask	0.5603	0.5134
Both	0.5538	0.5079

Table 5: Interaction–utility operating points. Matched-response controls are retrospective comparators using FITTA's realized response count; the deterministic control also matches its number of fine intervals. Random open-loop results summarize 20 pre-specified schedules. Bold and underlined values denote the best and second-best results, respectively, among operating points with 4.6145 responses/1k.

Operating point	Point AUPRC	VUS-PR	Responses / 1k
Fixed $H = 20$	0.5442	0.4978	4.1606
Deterministic matched budget	0.5414	0.4949	4.6145
Random open loop (mean)	0.5436	0.4970	4.6145
Random open loop (best)	<u>0.5489</u>	<u>0.5012</u>	4.6145
FITTA	**0.5511**	**0.5063**	4.6145
Fixed $H = 10$	0.5653	0.5224	8.3301

provides only modest additional aggregate utility, and its two uses are not additive. The aggregate effects are heterogeneous rather than uniformly positive: the normal-mask arm improves four of seven family means in Point AUPRC and three of seven in VUS-PR. It also introduces 28,530 additional masked-normal updates inside positive intervals, confirming that this oracle exposes adaptation opportunities unavailable to standard FITTA.

The utilization diagnostics help explain this ordering. About 35.8% of positive intervals contain normal timestamps that become available for additional catch-up under the normal-point mask, unlocking 327,921 normal timestamps in total. By contrast, FITTA's original source witness already falls on a true anomaly timestamp in 89.6% of positive intervals, leaving less localization headroom for the witness-only oracle. These oracle variants are diagnostic interventions rather than an upper bound on methods designed directly for point-level supervision. Full family-, seed-, and utilization-level results are reported in Appendix I.

5.5 Interaction–Utility Trade-off

Table 5 compares detection utility at different interaction rates. Fixed $H = 10$ reaches higher utility but requires 8.3301 responses/1k, versus 4.1606 for fixed $H = 20$ and 4.6145 for FITTA. At exactly FITTA's response count, FITTA has the highest equal-family macro utility among the deterministic matched-budget control and all 20 pre-specified random open-loop schedules on both metrics. Against the deterministic matched-budget control, its paired Point-AUPRC contrasts across seeds 42/43/44 are $+0.0002/+0.0100/+0.0191$. Relative to the random-control mean, FITTA is also higher in six of seven families on both metrics, with PUMP as the exception. The aggregate advantage is therefore not explained by response count alone. However, these controls do not isolate temporal placement as the sole cause, because different schedules also change returned feedback bits, confirmed-normal coverage, update timing, and subsequent optimization trajectories. Control details are in Appendix C.

5.6 Feedback-Error Sensitivity

Table 6 summarizes degradation from the clean reference under directional feedback corruption. Across all tested nonzero rates, $P \to N$ errors cause substantially larger aggregate degradation than $N \to P$ errors for both FITTA and fixed $H = 20$.

Table 6: Degradation from the clean reference under directional feedback corruption. Each entry reports Point AUPRC / VUS-PR degradation; negative values indicate a small improvement.

Arm	Direction	5%	10%	20%
FITTA	$P \to N$	0.0745 / 0.0672	0.1067 / 0.0986	0.1433 / 0.1327
FITTA	$N \to P$	-0.0015 / -0.0003	0.0047 / 0.0056	0.0209 / 0.0219
Fixed $H = 20$	$P \to N$	0.0639 / 0.0546	0.1007 / 0.0893	0.1325 / 0.1205
Fixed $H = 20$	$N \to P$	0.0048 / 0.0049	0.0070 / 0.0075	0.0207 / 0.0212

At 10%, FITTA degrades by 0.1067 Point AUPRC and 0.0986 VUS-PR under $\mathrm{P} \to \mathrm{N}$, versus 0.0047 and 0.0056 under $\mathrm{N} \to \mathrm{P}$. $\mathrm{P} \to \mathrm{N}$ degradation increases with corruption rate, whereas FITTA improves slightly under 5% $\mathrm{N} \to \mathrm{P}$ corruption; the directional ordering is not universal at family level. This ordering is consistent with the asymmetric update rules: $\mathrm{P} \to \mathrm{N}$ directs an anomaly-containing interval into confirmed-normal catch-up, whereas $\mathrm{N} \to \mathrm{P}$ suppresses normal catch-up and may invoke positive-evidence preservation; under adaptive FITTA, both error types can also alter subsequent interval selection. These results establish asymmetric directional sensitivity under the tested synthetic corruption process. Appendix E reports the shared clean references and full dual-metric results.

6 Limitations

Our FITTA evaluation spans seven benchmark families and two reconstruction-based backbones, so the observed gains should not be assumed to generalize to all TSAD settings or detector paradigms. Feedback is simulated from benchmark annotations rather than collected from human operators; the responses-per-1,000-timestamps metric therefore measures interaction frequency rather than annotation time, cognitive effort, or equal information cost. Adaptive temporal resolution also assumes that the feedback provider can support both tested interval lengths.

Our analysis is limited to the tested update objectives, temporal resolutions, and scheduling controls. The effects of preservation and adaptive resolution are condition-dependent, while the matched-response comparisons do not establish scheduler optimality or isolate temporal placement as a causal mechanism. The mechanism diagnostics are retrospective, and the corruption study uses synthetic rather than calibrated human-error models; these analyses therefore support descriptive mechanism and sensitivity conclusions rather than universal causal explanations or generic robustness to feedback errors. The point-localization oracle analysis evaluates two specific uses of exact point labels under the fixed FITTA update objectives and should not be interpreted as an upper bound on methods designed directly for point-level supervision.

7 Conclusion

We studied test-time adaptation for streaming time-series anomaly detection under coarse existential interval feedback, where each returned bit summarizes a completed interval without revealing anomaly locations and can influence only subsequent detection units. We introduced FITTA, a minimal adaptation method for this setting that uses the asymmetric information carried by negative and positive responses and adjusts the temporal resolution of subsequent feedback intervals.

Across seven benchmark families, FITTA improves equal-family Point AUPRC from 0.327 to 0.551 and VUS-PR from 0.248 to 0.506 relative to the frozen detector. Matched factorial analysis shows that most of this observed utility comes from adaptation on intervals certified as normal, while positive-evidence preservation and adaptive resolution provide a smaller, interacting refinement. Matched point-localization diagnostics yield only modest additional aggregate utility under the tested update objectives, while matched-response controls show that response count alone does not explain the adaptive result. Directional corruption tests further show substantially greater aggregate sensitivity to $P \to N$ than to $N \to P$ errors under the tested synthetic corruption process. The main utility gain also transfers to a second reconstruction-based detector. Overall, these results show that temporally coarse feedback can provide substantial adaptation utility for subsequent detection without requiring anomaly localization.

AI USE STATEMENT

Generative AI tools were used in this work to assist with research brainstorming and hypothesis refinement, refinement of experimental methodology, implementation and code-related tasks, interpretation and auditing of experimental results, and literature retrieval and synthesis. They were also used to assist with drafting and editing portions of the manuscript for structure, clarity, and readability. We did not use generative AI to generate synthetic experimental datasets or to establish mathematical proofs. AI-assisted research outputs underlying the reported experimental results and claims were subject to author verification through experimental execution, protocol and implementation checks, and comparison against the underlying experimental artifacts. AI-assisted manuscript changes were reviewed by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the implementation and evaluation specifications needed to reproduce the reported experimental protocol throughout the main paper and appendix. Section 5.1 and Appendix B specify the evaluation scope, preprocessing, windowing, model architectures, source-training and test-time optimization settings, metrics, aggregation procedure, and random seeds. Appendix A gives the exact chronological feedback and adaptation procedure, including confirmed-normal catch-up, positive-evidence preservation, optimizer-state handling, terminal-interval semantics, and operational assumptions. Appendices C–I document the matched-response controls, preservation analysis across temporal resolutions, feedback-corruption protocol, cross-backbone evaluation, mechanism diagnostics, family- and seed-level breakdowns, and point-localization oracle diagnostics. The implementation and evaluation code will be released upon publication.

REFERENCES

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: unsupervised anomaly detection on multivariate time series. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 3395–3404. ACM, 2020. doi: 10.1145/3394486.3403392. URL https://doi.org/10.1145/3394486.3403392.

Ane Blázquez-García, Angel Conde, Usue Mori, and José Antonio Lozano. A review on outlier/anomaly detection in time series data. *ACM Comput. Surv.*, 54(3):56:1–56:33, 2022.

Hamza Bodor, Thai V. Hoang, and Zonghua Zhang. Little help makes a big difference: Leveraging active learning to improve unsupervised time series anomaly detection. In Hakim Hacid, Monther Aldwairi, Mohamed Reda Bouadjenek, Marinella Petrocchi, Noura Faci, Fatma Outay, Amin Beheshti, Lauritz Thamsen, and Hai Dong (eds.), *Service-Oriented Computing – ICSOC 2021 Workshops – AIOps, STRAPS, AI-PA and Satellite Events, Dubai, United Arab Emirates, November 22–25, 2021, Proceedings*, volume 13236 of *Lecture Notes in Computer Science*, pp. 165–176. Springer, 2022. doi: 10.1007/978-3-031-14135-5_13. URL https://doi.org/10.1007/978-3-031-14135-5_13.

Andrew A. Cook, Goksel Misirli, and Zhong Fan. Anomaly detection for IoT time-series data: A survey. *IEEE Internet of Things Journal*, 7(7):6481–6494, 2020.

Hong Guo, Yujing Wang, Jieyu Zhang, Zhengjie Lin, Yunhai Tong, Lei Yang, Luoxing Xiong, and Congrui Huang. Label-efficient interactive time-series anomaly detection. *CoRR*, abs/2212.14621, 2022. doi: 10.48550/ARXIV.2212.14621. URL https://doi.org/10.48550/ARXIV.2212.14621.

Seung Hun Han, Hyeongwon Kang, Jinwoo Park, and Pilsung Kang. Detecting the undetectable: Enhancing unsupervised time series anomaly detection via active learning. *CoRR*, abs/2607.00720, 2026. doi: 10.48550/ARXIV.2607.00720. URL https://doi.org/10.48550/ARXIV.2607.00720.

Tao Huang, Pengfei Chen, and Ruipeng Li. A semi-supervised VAE based active anomaly detection framework in multivariate time series for online systems. In *The ACM Web Conference 2022*, pp. 1797–1806. ACM, 2022. doi: 10.1145/3485447.3511984. URL https://doi.org/10.1145/3485447.3511984.

Dongmin Kim, Sunghyun Park, and Jaegul Choo. When model meets new normals: Test-time adaptation for unsupervised time-series anomaly detection. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence*, pp. 13113–13121. AAAI Press, 2024. doi: 10.1609/AAAI.V38I12.29210. URL https://doi.org/10.1609/AAAI.V38I12.29210.

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 Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence*, pp. 15018–15026. AAAI Press, 2026. doi: 10.1609/AAAI.V40I17.38524. URL https://doi.org/10.1609/AAAI.V40I17.38524.

Dongha Lee, Sehun Yu, Hyunjun Ju, and Hwanjo Yu. Weakly supervised temporal anomaly segmentation with dynamic time warping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7355–7364, 2021. doi: 10.1109/ICCV48922.2021.00726. URL https://doi.org/10.1109/ICCV48922.2021.00726.

Chen Liu, Shibo He, Haoyu Liu, and Shizhong Li. TreeMIL: A multi-instance learning framework for time series anomaly detection with inexact supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7510–7514. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447536. URL https://doi.org/10.1109/ICASSP48485.2024.10447536.

Dapeng Liu, Youjian Zhao, Haowen Xu, Yongqian Sun, Dan Pei, Jiao Luo, Xiaowei Jing, and Mei Feng. Opprentice: Towards practical and automatic anomaly detection through machine learning. In Kenjiro Cho, Kensuke Fukuda, Vivek S. Pai, and Neil Spring (eds.), *Proceedings of the 2015 ACM Internet Measurement Conference, IMC 2015, Tokyo, Japan, October 28–30, 2015*, pp. 211–224. ACM, 2015. doi: 10.1145/2815675.2815679. URL https://doi.org/10.1145/2815675.2815679.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Wenlu Wang, Pengfei Chen, Yibin Xu, and Zilong He. Active-MTSAD: Multivariate time series anomaly detection with active learning. In *52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 263–274. IEEE, 2022. doi: 10.1109/DSN53405.2022.00036. URL https://doi.org/10.1109/DSN53405.2022.00036.

Yaxuan Wang, Hao Cheng, Jing Xiong, Qingsong Wen, Han Jia, Ruixuan Song, Liyuan Zhang, Zhaowei Zhu, and Yang Liu. Noise-resilient point-wise anomaly detection in time series using weak segment labels. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1551–1562. ACM, 2025a. doi: 10.1145/3690624.3709257. URL https://doi.org/10.1145/3690624.3709257.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *Int. J. Comput. Vis.*, 133(3):1106–1139, 2025b. doi: 10.1007/S11263-024-02213-5. URL https://doi.org/10.1007/s11263-024-02213-5.

Qingsong Wen, Linxiao Yang, Tian Zhou, and Liang Sun. Robust time series analysis and applications: An industrial perspective. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 4836–4837, 2022.

Xu Zhang, Qingwei Lin, Yong Xu, Si Qin, Hongyu Zhang, Bo Qiao, Yingnong Dang, Xinsheng Yang, Qian Cheng, Murali Chintalapati, Youjiang Wu, Ken Hsieh, Kaixin Sui, Xin Meng, Yaohai

Xu, Wenchi Zhang, Furao Shen, and Dongmei Zhang. Cross-dataset time series anomaly detection for cloud systems. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, pp. 1063–1076. USENIX Association, 2019.

Xiaohui Zhou, Yijie Wang, Hongzuo Xu, and Mingyu Liu. Boundary-driven active learning for anomaly detection in time series data streams. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6135–6139. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446524. URL https://doi.org/10.1109/ICASSP48485.2024.10446524.

A Detailed FITTA Online Operation

For each physical stream, FITTA maintains an immutable source reference θ_0 and a continually adapted deployment state θ. All detection units in the current feedback interval are processed chronologically and their anomaly scores are committed before interval feedback is returned. No parameter update occurs while the interval is being scored.

Confirmed-normal catch-up. After a negative response, the detection units in the completed interval are re-forwarded once in their original chronological order. One optimization step is performed per unit using $\mathcal{L}_{\mathrm{N}}$, and each later step uses the parameters produced by the preceding catch-up step. There is no interval-wide averaged update and no replay of earlier feedback intervals. These computations are used only for adaptation and never replace committed scores.

Positive-evidence preservation. After a positive response, normal catch-up is disabled for the entire interval. The completed interval is re-forwarded under the current deployment model to construct a differentiable graph for $\mathcal{L}_{\mathrm{P}}$. The immutable source detector supplies $S_0(I_k)$ and the deployment model supplies $S_\theta(I_k)$; their maximizing timestamps need not coincide. If $S_\theta(I_k) \geq S_0(I_k)$, no preservation step is taken. Otherwise, exactly one step penalizes attenuation of source-relative interval-maximum evidence. It is not required to restore the inequality after one update.

Persistent state and terminal intervals. Both branches update the same deployment model and share optimizer state within a physical deployment. The returned bit determines the temporal resolution of the next interval. A terminal incomplete interval is scored and evaluated but receives no feedback and triggers no adaptation. No anomaly location or event boundary is available to the deployed algorithm.

Operational overhead. FITTA retains an immutable copy of the source detector in addition to the adapted deployment model and buffers the completed detection units of the current feedback interval until its response is returned. Optimizer state is retained throughout each deployment. A confirmed-normal interval triggers one sequential adaptation step per completed detection unit, so adaptation work is bursty at feedback boundaries rather than uniformly distributed over the stream. Our chronological evaluation establishes that feedback-dependent updates affect only subsequent detection units, but does not by itself establish real-time throughput or latency guarantees.

B Additional Experimental and Protocol Details

Evaluation scope. Our evaluation comprises GHL, CATSv2, PUMP, SMD, PSM, SWAT, and GECCO. Each entity is evaluated independently with seeds 42, 43, and 44. This fixed scope is not a random or representative sample of TSAD benchmarks.

Primary autoencoder and source training. For an entity with C retained channels, each standardized length-12 window is flattened to dimension $D = 12C$. The primary autoencoder uses a three-layer encoder

$$D \to h_1 \to h_2 \to z,$$

where

$$h_1 = \max(1, \lfloor D/2 \rfloor), \qquad h_2 = \max(1, \lfloor D/4 \rfloor), \qquad z = \min(64, \max(8, \lfloor D/8 \rfloor)).$$

Each encoder Linear layer is followed by ReLU. The decoder mirrors these dimensions as

$$z \to h_2 \to h_1 \to D,$$

with ReLU after its first two Linear layers and a linear output layer. The model uses no dropout or internal normalization and reconstructs the standardized input window.

Source data are split chronologically at 80/20 before window construction, so no window crosses the train–validation boundary. Source training uses Adam with learning rate 10^{-4}, zero weight decay, batch size 64, and at most 10 epochs. The first two epochs use learning rate 10^{-4}; from the third epoch onward it is halved at each epoch. Training stops after three consecutive epochs without

a strict decrease in validation loss, and the model state with the lowest validation loss is restored. Validation loss is the unweighted mean of the per-batch reconstruction-MSE means. Separate source checkpoints are trained for seeds 42, 43, and 44.

Normalization. Normalization is performed independently for each retained channel of each physical entity. Let μ_c and σ_c denote the mean and population standard deviation (`ddof`=0) of channel c over the complete source partition. We apply

$$\widetilde{x}_{r,c} = \frac{x_{r,c} - \mu_c}{a_c}, \qquad a_c = \begin{cases} 1, & \text{if channel } c \text{ is constant or } \sigma_c = 0, \\ \sigma_c, & \text{otherwise.} \end{cases}$$

The statistics are estimated before the chronological 80/20 source train–validation split and are reused unchanged for deployment data. Statistics are computed in float64 and transformed arrays are stored in float32. No additive epsilon, clipping, or online refitting is used.

Deployment chronology and optimization. Detection uses non-overlapping length-12, stride-12 windows and online batch size one. FITTA uses $(H_{\text{coarse}}, H_{\text{fine}}) = (20, 10)$ and AdamW with learning rate 10^{-4} and zero weight decay. Optimizer state persists within each physical deployment and is reset between deployments. Every complete window is evaluated; raw observations that cannot form a complete window are excluded.

Metrics and aggregation. Point AUPRC is `auc(recall, precision)` applied to the arrays returned by `precision_recall_curve(labels, scores)`. It is neither average precision nor a point-adjusted or event-adjusted metric.

VUS-PR is computed with the local `range_auc_volume(labels, scores, window_size=12)` implementation on the same committed chronological timestamp-level scores and labels. We average the PR areas over the 13 integer tolerance settings $w = 0, \ldots, 12$, using 250 rank-sampled score thresholds. These tolerance values are measured in raw timestamp-index positions, not detection windows or wall-clock time. Under the implementation's endpoint convention, $w = 12$ extends a positive range by up to six preceding and five following timestamp positions.

Both metrics are computed independently for each physical entity and seed, averaged equally across entities within each family and seed, then across seeds, and finally across the seven families with equal weight. Timestamps are never pooled across entities.

For each physical entity and deployment seed, if K responses are returned over N_{eval} evaluated raw timestamps, interaction frequency is $1000K/N_{\text{eval}}$. These entity-level rates are averaged over entities within each family–seed condition, then over seeds, and finally over the seven families with equal family weight. This is the aggregation convention used for the reported responses-per-1,000 values. It measures returned-bit frequency, not annotation time or equal information content across feedback primitives.

Frozen, M2N2, and CANDI share checkpoints, preprocessing, windowing, and support where applicable while retaining method-specific adaptation. M2N2 and CANDI receive no external feedback. BALAD retains its native detector/protocol and point-specific endpoint feedback; exact BALAD–FITTA common support is used where applicable. It is therefore not a strictly matched protocol comparison.

For CANDI, we retain its dual-sided SANA modules, hard and moderate false-positive mining, separate accumulation buffers, frozen backbone, and persistent SGD, while adapting the method to the common chronological evaluation protocol. Thus, the CANDI row uses the source checkpoint, preprocessing, non-overlapping length-12 windows, chronological arrivals, and evaluation support rather than reproducing CANDI's native experimental pipeline. The harmonized implementation uses SANA width 512 with eight attention heads, feed-forward width 512, one layer, zero dropout, and zero-initialized gates. Adaptation uses persistent SGD with learning rate 10^{-4}, momentum 0.9, weight decay 10^{-4}, and one optimization step whenever either the hard or moderate buffer accumulates 16 samples. False-positive mining uses the 95th percentile of the source-validation scores ($\alpha = 5\%$) as the score threshold, together with the reference Mahalanobis criterion. The released CANDI configurations are dataset- and entity-specific and do not cover all seven evaluated

benchmark families, so we do not retune its adaptation hyperparameters using test performance on the evaluated benchmarks. In the reported runs, CANDI performed adaptation updates in 97.2% of evaluated entity–seed deployments; its full-precision aggregate metrics differ slightly from Frozen but coincide after three-decimal rounding.

C Full Interaction–Utility Results

The matched-response analysis uses the same seven-family evaluation scope and deployment seeds 42, 43, and 44.

Deterministic matched-budget comparator. For each physical entity and deployment seed, let K denote the number of completed feedback intervals in the paired adaptive FITTA trajectory and let m denote the number of those intervals that use $H = 10$. The deterministic control matches both quantities. When $m > 0$, among its K response slots, $H = 10$ is placed at the zero-based indices

$$i_j = \left\lfloor \frac{(2j+1)K}{2m} \right\rfloor, \qquad j = 0, \ldots, m-1,$$

with all remaining slots using $H = 20$. When $m = 0$, all K slots use $H = 20$. Interval boundaries are the cumulative sums of these lengths. Any remaining tail shorter than 20 detection windows is scored but receives no response and triggers no adaptation. Once constructed, the control schedule is fixed and does not respond to its own feedback. Because K and m are obtained from the completed paired FITTA trajectory, this is a retrospective matched-budget comparator rather than a prospectively budget-known online policy.

Random open-loop controls. We additionally evaluate 20 pre-specified feedback-independent open-loop schedules with schedule seeds 1001–1020. Separately for every entity–deployment-seed pair, each schedule returns exactly the realized response count of the paired FITTA trajectory. Given the stream length and this count, the construction uniformly selects an admissible number of $H = 10$ intervals, fills the remaining responses with $H = 20$ intervals, and randomly permutes the resulting sequence, keyed by the schedule seed and immutable deployment identity. The matched response count is taken retrospectively from the paired FITTA trajectory; conditional on this count, response placement is independent of labels, scores, feedback values, and FITTA interval locations.

Source checkpoints, AE architecture, preprocessing, length-12/stride-12 evaluation support, the adaptation update rules, optimizer, online batch size, the requirement that each window is scored before feedback or adaptation, and response count are matched.

Table 7: Matched-response scheduling controls across the seven evaluated benchmark families. Random open-loop results summarize 20 pre-specified feedback-independent schedules.

Arm	Point AUPRC	VUS-PR	Responses/1k
FITTA	0.5511	0.5063	4.6145
Deterministic matched budget	0.5414	0.4949	4.6145
Random open loop (mean)	0.5436	0.4970	4.6145
Random open loop (best)	0.5489	0.5012	4.6145

At the equal-family macro level, FITTA exceeds the deterministic control and all 20 random schedules on both metrics. Relative to the random-control mean, FITTA is higher in six of seven families on both metrics, with PUMP as the exception. The individual random-schedule aggregates are reported in Table 8.

These controls show that response count alone is insufficient to explain FITTA's aggregate advantage. They cover only a finite set of open-loop policies and therefore do not establish scheduler optimality or isolate temporal placement as the sole causal mechanism.

Table 8: Per-schedule equal-family macro results for the 20 pre-specified feedback-independent open-loop schedules. Every row has 4.6145 responses/1k.

Seed	Point AUPRC	VUS-PR	Seed	Point AUPRC	VUS-PR
1001	0.5433	0.4960	1011	0.5405	0.4946
1002	0.5423	0.4967	1012	0.5393	0.4938
1003	0.5436	0.4963	1013	0.5470	0.5003
1004	0.5489	0.5012	1014	0.5425	0.4965
1005	0.5471	0.5000	1015	0.5447	0.4978
1006	0.5475	0.4990	1016	0.5429	0.4954
1007	0.5422	0.4961	1017	0.5407	0.4955
1008	0.5451	0.4988	1018	0.5385	0.4938
1009	0.5411	0.4946	1019	0.5434	0.4959
1010	0.5457	0.4984	1020	0.5456	0.4986

D Positive-Evidence Preservation Across Temporal Resolution

This analysis fixes $H \in \{5, 10, 20, 40, 80\}$ and disables adaptive switching. Preservation OFF uses confirmed-normal catch-up only; Preservation ON uses the same adaptation procedure plus the Positive-Evidence Preservation update. Within each H, temporal resolution, response opportunities, source checkpoints, AE architecture, preprocessing, optimizer, confirmed-normal update rule, length-12/stride-12 evaluation support, online batch size, and the requirement that each window is scored before feedback or adaptation are matched. The reported delta is Preservation ON minus Preservation OFF.

Table 9: Incremental effect of Positive-Evidence Preservation across fixed temporal resolutions over the seven evaluated benchmark families. Family and seed columns report positive/negative preservation effects. Deltas are computed from unrounded values.

H	Metric	OFF	ON	Δ	Families $+/-$	Seeds $+/-$
5	Point AUPRC	0.5602	0.5946	+0.0344	6/1	3/0
	VUS-PR	0.5149	0.5519	+0.0370	6/1	3/0
10	Point AUPRC	0.5481	0.5653	+0.0171	4/3	3/0
	VUS-PR	0.5027	0.5224	+0.0198	5/2	3/0
20	Point AUPRC	0.5411	0.5442	+0.0032	5/2	2/1
	VUS-PR	0.4928	0.4978	+0.0050	5/2	2/1
40	Point AUPRC	0.5097	0.5074	−0.0023	3/4	1/2
	VUS-PR	0.4619	0.4607	−0.0011	3/4	1/2
80	Point AUPRC	0.4792	0.4761	−0.0031	3/4	1/2
	VUS-PR	0.4302	0.4279	−0.0023	2/5	1/2

The aggregate preservation effect decreases across the five tested temporal resolutions for both metrics and changes sign between $H = 20$ and $H = 40$. It is positive in aggregate at $H = 5$–20 and slightly negative at $H = 40$–80. Family-level effects are heterogeneous at every resolution. These results characterize resolution dependence but do not identify the mechanism behind the sign change or establish an optimal temporal resolution.

E Feedback-Error Sensitivity

For P→N corruption, a positive returned bit is flipped to negative; for N→P, a negative bit is flipped to positive. Rates are 0%, 5%, 10%, and 20% among eligible responses, while clean point labels remain fixed for evaluation. The 0% condition is one shared clean reference per arm; its use in both directional contrasts is not independent clean evidence. For FITTA, the corrupted returned bit affects both adaptation and the controller state that selects the next interval. For fixed $H = 20$, corruption affects adaptation but cannot change the temporal resolution.

Eligible responses are complete intervals whose clean existential bit is positive for $P \to N$ and negative for $N \to P$. Each eligible response is flipped by a deterministic Bernoulli-style hash test at

the prescribed rate q, using SHA-256 of the run identity together with the interval index and start/end base-window indices; the procedure does not select an exact fraction of eligible responses. Run identities distinguish entities, deployment seeds, arms, and corruption directions, and we use one realization per condition for each entity and deployment seed. For adaptive FITTA, corruption plans are constructed sequentially along the corrupted-feedback trajectory, so a flipped bit can change subsequent interval boundaries, eligibility, and later corruption decisions.

Table 10: Point AUPRC under directional feedback corruption. Parentheses give degradation from the shared arm-specific clean reference; negative degradation denotes a small improvement. Degradations are computed from unrounded values.

Arm	Direction	Clean	5%	10%	20%
FITTA	P→N	0.5511	0.4767 (0.0745)	0.4444 (0.1067)	0.4079 (0.1433)
FITTA	N→P	0.5511	0.5526 (−0.0015)	0.5464 (0.0047)	0.5302 (0.0209)
Fixed $H = 20$	P→N	0.5442	0.4804 (0.0639)	0.4435 (0.1007)	0.4118 (0.1325)
Fixed $H = 20$	N→P	0.5442	0.5394 (0.0048)	0.5372 (0.0070)	0.5235 (0.0207)

Table 11: VUS-PR under directional feedback corruption. Parentheses give degradation from the shared arm-specific clean reference; negative degradation denotes a small improvement. Degradations are computed from unrounded values.

Arm	Direction	Clean	5%	10%	20%
FITTA	P→N	0.5063	0.4391 (0.0672)	0.4077 (0.0986)	0.3736 (0.1327)
FITTA	N→P	0.5063	0.5066 (−0.0003)	0.5007 (0.0056)	0.4844 (0.0219)
Fixed $H = 20$	P→N	0.4978	0.4432 (0.0546)	0.4085 (0.0893)	0.3773 (0.1205)
Fixed $H = 20$	N→P	0.4978	0.4929 (0.0049)	0.4903 (0.0075)	0.4766 (0.0212)

Across all tested nonzero corruption rates, aggregate P→N degradation exceeds N→P degradation for both FITTA and fixed $H = 20$ on both metrics, and P→N degradation increases with corruption rate. FITTA improves slightly under 5% N→P corruption, so its N→P trajectory is not strictly monotonic from clean. The directional ordering is not universal at the family level. These experiments therefore establish directional sensitivity under the tested synthetic corruption process, not generic robustness or a calibrated model of human or operator annotation errors. Because corrupted feedback in FITTA affects both adaptation and subsequent interval selection, the results do not causally decompose updater and controller effects.

F Cross-Backbone Results

USAD supplies a second reconstruction-based backbone under the same seven-family evaluation scope and interval-feedback semantics. Absolute AE and USAD scores should not be interpreted as a detector ranking.

USAD instantiation. For an input window with flattened dimension $D = 12C$, USAD uses a shared encoder

$$D \to \max(1, \lfloor D/2 \rfloor) \to \max(1, \lfloor D/4 \rfloor) \to 5$$

and two separate mirrored decoders. ReLU follows each encoder layer and the first two layers of each decoder; the decoder outputs are linear.

Let

$$\hat{x}_\theta^{(1)} = D_1(E(x)), \qquad \hat{x}_\theta^{(2)} = D_2(E(\hat{x}_\theta^{(1)})),$$

and let

$$e(\hat{x}, x)_r = \frac{1}{C} \sum_{d=1}^{C} (\hat{x}_{r,d} - x_{r,d})^2.$$

The deployed timestamp-level anomaly score is

$$s_\theta(r) = \frac{1}{2} e(\hat{x}_\theta^{(1)}, x)_r + \frac{1}{2} e(\hat{x}_\theta^{(2)}, x)_r.$$

The direct reconstruction $D_2(E(x))$ is not included in this deployment score.

During FITTA adaptation, the shared encoder and both decoders are trainable. For a confirmed-normal interval, each window receives one chronological update minimizing the temporal mean of the composite score above. For a positive interval, FITTA uses the same source-relative preservation objective as in the main method,

$$\left[\max_{r \in \mathcal{R}(I_k)} s_{\theta_0}(r) - \max_{r \in \mathcal{R}(I_k)} s_\theta(r) \right]_+ ,$$

with the immutable source USAD supplying the reference score. The native epoch-dependent USAD source-training loss is not reused for test-time adaptation.

Deployment uses the same persistent AdamW configuration as the primary AE: learning rate 10^{-4}, betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, and zero weight decay. A fresh optimizer is initialized for each physical deployment and retained across its negative and positive feedback updates. As in the primary experiments, all interval scores are committed before feedback and adaptation, and incomplete terminal intervals trigger no update.

Table 12: USAD family-level cross-backbone results and equal-family macro. Entries are Point AUPRC / VUS-PR.

USAD arm	GHL	CATSv2	PUMP	SMD	PSM	SWAT	GECCO	Macro
Frozen	.014/.014	.123/.104	.248/.258	.441/.403	.536/.483	.715/.400	.203/.035	.326/.242
Fixed $H = 20$ + preservation	.074/.070	.263/.251	.718/.731	.476/.439	.672/.644	.831/.685	.685/.544	.531/.480
Deterministic matched budget	.068/.065	.268/.255	.719/.725	.482/.446	.672/.637	.838/.686	.684/.544	.533/.480
FITTA	.107/.100	.300/.288	.744/.748	.480/.443	.659/.631	.847/.702	.692/.563	.547/.496

FITTA improves over Frozen on this second reconstruction backbone and exceeds both tested controls in the macro. Family-level responses are heterogeneous, so the result supports transfer to this particular backbone rather than architecture invariance.

G Full Mechanism Diagnostics

All diagnostics in this section are retrospective. Hidden point labels are used to characterize saved intervals and witnesses only; they never enter the deployed update rule. Within each cadence, the preservation ON/OFF arms are paired on exact chronological scores, evaluation support, and feedback boundaries; fixed and adaptive cadence use different temporal partitions.

G.1 Partitioning and Positive-Region Concentration

Table 13: Partition and opportunity diagnostics. Raw-count rows report pooled totals, whereas rates and coverage are equal-family aggregates and therefore need not equal ratios computed from the raw totals.

Quantity	Fixed $H = 20$	Adaptive
Returned intervals (raw total)	73,032	76,290
Positive intervals (raw total)	4,200	6,480
Responses/1k	4.1606	4.6145
Positive-response fraction	0.1406	0.1825
Confirmed-normal coverage	0.8582	0.8664
Positive-interval anomaly occupancy	0.6380	0.7072
Active preservation intervals	2,626	3,506
Active preservation / 1k	0.3352	0.4339
Per-positive activation rate	0.6138	0.5551

Adaptive resolution increases, rather than reduces, confirmed-normal coverage and normal catch-up steps. Across 21 family–seed units, the association between its performance change and coverage change is weakly negative (Spearman $\rho = -0.252$), while the association with update-step change is near zero ($\rho = -0.043$). Reduced normal supervision is therefore not supported as the explanation for the adaptive normal-only decrement.

Recent positive feedback marks temporally concentrated regions: $P(Y_{k+1} = 1 \mid Y_k = 1)$ is 0.726 under fixed $H = 20$ and 0.739 under adaptive resolution, compared with 0.056 and 0.060 following a negative response. Adaptive positive intervals also have higher retrospective anomaly occupancy (0.707 versus 0.638). This is association, partly including continuation of the same event, rather than a causal or deployed forecast.

G.2 Activation, Witnesses, and Score Shifts

Adaptive resolution creates more active preservation opportunities in total and per timestamp, but not a higher activation probability conditional on a positive interval. The source-witness anomaly hit rate rises from 0.8571 to 0.9055 and the adapted-witness rate from 0.8626 to 0.9061. Conditional on activation, the corresponding source/adapted hit rates are 0.8153/0.8240 under fixed cadence and 0.8719/0.8740 under adaptive cadence. These results support witness validity in aggregate, not point localization by the deployed method.

For the matched preservation-versus-normal-only score comparison, the equal-family mean anomaly-score shift is -20.07 under fixed cadence and -32.92 under adaptive cadence; mean normal-score shifts are -0.153 and -0.099. Median anomaly shifts are -0.922 and $+0.538$, and the fractions of anomaly points with positive shifts are 0.555 and 0.588. In the immediately following interval, mean anomaly shifts are -54.03 and -1262.79, whereas median shifts are $+4.21$ and -1371.01 for fixed and adaptive cadence, respectively. The magnitude and direction are highly heterogeneous across families and score scales. These shift magnitudes are expressed in detector-specific raw score units and are therefore not directly comparable across families. These diagnostics do not support a universal claim that preservation prevents future anomaly-score attenuation.

For the family-level diagnostic below, we define the preservation-induced separation shift for each entity–seed deployment under adaptive resolution as

$$\Delta_{\text{sep}} = \mathbb{E}_{y_r=1}\left[s_r^{\text{N+P}} - s_r^{\text{N}}\right] - \mathbb{E}_{y_r=0}\left[s_r^{\text{N+P}} - s_r^{\text{N}}\right],$$

where $\text{N} + \text{P}$ denotes confirmed-normal adaptation with Positive-Evidence Preservation and N denotes the matched confirmed-normal-only condition. Positive values indicate increased mean anomaly–normal score separation under preservation, whereas negative values indicate reduced separation. The quantity is expressed in unnormalized detector-score units and its magnitude is therefore not directly comparable across families. The table reports equal-entity averages within each family and seed, followed by an equal-seed family mean.

Table 14: Family heterogeneity in selected mechanism diagnostics. Coverage and activation columns are adaptive-minus-fixed changes; witness hit rate and separation shift are reported for adaptive resolution.

Family	Δ normal coverage	Δ activation rate	Source hit	Separation shift
GHL	+.0004	-.1526	.812	+.278
CATSv2	+.0033	-.0114	.921	+.360
PUMP	+.0050	-.0638	.967	-.439
SMD	+.0116	-.0102	.890	-244.542
PSM	+.0137	-.1135	.901	+.272
SWAT	+.0061	-.0790	.932	+2.077
GECCO	+.0177	+.0196	.915	+12.240

Mechanism verdict. The evidence supports increased and more concentrated positive-feedback opportunities and generally valid interval-maximum witnesses. It does not support reduced normal supervision as the cause of the normal-only decrement, nor universal future attenuation prevention. The opportunity mechanism is therefore only partially supported as an explanation of the positive interaction between preservation and temporal resolution; alternative trajectory-level explanations remain unresolved.

H FAMILY AND SEED BREAKDOWN

Table 15: Matched component AUPRC contrasts. Family signs count positive/negative family means; seed signs follow seeds 42/43/44.

Contrast	Macro Δ	Families $+/-$	Seed signs
Fixed normal-only − Frozen	+0.2144	7/0	+/+/+
Preservation effect, fixed	+0.0032	5/2	-/+/+
Preservation effect, adaptive	+0.0164	5/2	+/+/+
Adaptive effect, normal-only	-0.0063	3/4	-/-/-
Adaptive effect, + preservation	+0.0069	6/1	-/+/+

Table 16: Family-level Point-AUPRC component contrasts. Columns report, respectively, preservation under fixed cadence, preservation under adaptive cadence, adaptive resolution without preservation, and adaptive resolution with preservation.

Family	Pres./fixed	Pres./adaptive	Adaptive/no pres.	Adaptive/pres.
GHL	+.0069	+.0257	-.0015	+.0173
CATSv2	+.0001	+.0098	+.0100	+.0197
PUMP	-.0219	-.0103	-.0275	-.0159
SMD	+.0017	+.0002	+.0027	+.0012
PSM	+.0350	+.0731	-.0304	+.0077
SWAT	+.0098	+.0166	+.0047	+.0115
GECCO	-.0094	-.0003	-.0023	+.0068

The family table exposes rather than hides contradictory cases: PUMP is negative for all four component contrasts, while the positive adaptive effect with preservation holds for six of seven family means. Seed signs likewise show that the adaptive-versus-fixed contrast with preservation enabled is negative for seed 42 and positive for seeds 43 and 44.

Across seeds 42/43/44, FITTA–Frozen is $+0.2195/+0.2202/+0.2337$ for Point AUPRC and $+0.2554/+0.2575/+0.2630$ for VUS-PR. The VUS-PR preservation contrasts are $-0.0005/+0.0066/+0.0089$ at fixed $H = 20$ and $+0.0078/+0.0288/+0.0211$ under adaptive resolution. FITTA minus the deterministic matched-budget control is $+0.0029/+0.0131/+0.0182$ for VUS-PR. These are descriptive differences across three fixed initialization/optimization seeds on the same seven-family evaluation panel, rather than independent benchmark replications; entity–seed observations are not treated as independent datasets for significance claims.

I POINT-LOCALIZATION ORACLE DIAGNOSTICS

We further examine how much additional adaptation utility becomes available when exact point-level localization is revealed after a positive feedback interval has already been fully scored. Point labels are exposed to the updater only in this oracle analysis and are never available before scoring. All committed scores therefore remain unchanged, and oracle information can affect future updates only.

The analysis uses the same seven benchmark families, deployment seeds 42/43/44, source checkpoints, preprocessing, primary AE, optimizer, and adaptive feedback schedule as the main FITTA evaluation. The interval controller continues to depend only on the existential response Y_{I_k}, so the four arms have identical feedback-interval boundaries, response counts, and evaluation support. Across the matched entity–seed deployments, every arm returns the same pooled total of 76,290 responses over the same 17,545,464 evaluated raw timestamps. The reported interaction rate of 4.6145 responses per 1,000 follows the entity-to-family aggregation convention defined in Appendix B.

Oracle variants. The reference arm is standard FITTA, which receives no point-level localization. We denote standard FITTA, witness only, normal mask only, and both by A, B, C, and D, respectively. The three oracle variants expose exact point labels only after a positive interval has been completed.

Table 17: Matched point-localization oracle results. Deltas are relative to standard FITTA and are computed from the audited unrounded aggregates. All four arms have identical feedback schedules, response counts, and evaluation support.

Localization	Point AUPRC	Δ	VUS-PR	Δ
None (FITTA)	0.5511	–	0.5063	–
Witness only	0.5547	+0.0036	0.5111	+0.0048
Normal mask only	0.5603	+0.0091	0.5134	+0.0071
Both	0.5538	+0.0026	0.5079	+0.0016

Witness only. Confirmed-normal catch-up remains disabled for a positive interval, but the preservation witness is restricted to timestamps that are truly anomalous:

$$S_\theta^+(I_k) = \max_{r \in \mathcal{R}(I_k):\, y_r = 1} s_\theta(r), \qquad S_0^+(I_k) = \max_{r \in \mathcal{R}(I_k):\, y_r = 1} s_{\theta_0}(r),$$

with

$$\mathcal{L}_P^+(I_k) = \left[S_0^+(I_k) - S_\theta^+(I_k) \right]_+ .$$

Normal mask only. The original interval-wide preservation objective is retained. In addition, point labels reveal which timestamps inside the positive interval are actually normal. For each completed detection unit X_t, define

$$\mathcal{N}_t = \{ r \in \mathcal{R}(X_t) : y_r = 0 \} .$$

If $\mathcal{N}_t$ is nonempty, one chronological adaptation step minimizes

$$\mathcal{L}_N^{\text{mask}}(X_t; \theta) = \frac{1}{|\mathcal{N}_t|} \sum_{r \in \mathcal{N}_t} s_\theta(r).$$

Detection units containing no normal timestamps are skipped. In both mask-enabled oracle arms (C and D), masked-normal updates are applied first in chronological order. The preservation score is then recomputed under the resulting deployment state and its gate is evaluated once, using the same persistent AdamW state and fixed source reference.

Both. This arm uses the same mask-first sequence as the normal-mask oracle, but replaces interval-wide preservation with the anomaly-restricted witness used by the witness-only oracle.

The normal-mask-only oracle provides the largest individual aggregate gain. Combining the two oracle uses does not preserve that gain: the descriptive 2×2 interaction,

$$D - B - C + A,$$

is -0.0100 for Point AUPRC and -0.0103 for VUS-PR. Thus, under the tested update objectives, the two uses of exact localization interact non-additively rather than producing an additive improvement.

Family and seed heterogeneity. Table 18 reports family-level Point-AUPRC contrasts relative to FITTA. All three oracle variants improve four of seven family means and reduce three. In particular, the larger normal-mask-only aggregate gain is driven mainly by PUMP, PSM, and GECCO, while GHL decreases substantially. This heterogeneity cautions against interpreting the aggregate oracle gain as a uniform benefit of point localization.

The corresponding VUS-PR contrasts are shown in Table 19. The qualitative picture is similarly heterogeneous: witness-only is positive on four of seven families, whereas normal-mask-only and the combined oracle are positive on three of seven.

Across seeds 42/43/44, the Point-AUPRC contrasts relative to FITTA are $+0.0126/+0.0019/-0.0039$ for witness-only, $+0.0138/+0.0075/+0.0061$ for normal-mask-only, and $+0.0114/+0.0016/-0.0050$ for the combined oracle. The corresponding VUS-PR contrasts are $+0.0146/+0.0017/-0.0020$, $+0.0145/+0.0038/+0.0032$, and $+0.0127/-0.0024/-0.0056$, respectively. These are descriptive contrasts across three fixed deployment seeds rather than independent benchmark replications.

Table 18: Family-level Point-AUPRC contrasts relative to FITTA for the point-localization oracle variants.

Family	Witness	Normal mask	Both
GHL	-0.0021	-0.0192	-0.0185
CATSv2	+0.0045	-0.0029	-0.0009
PUMP	+0.0172	+0.0358	+0.0193
SMD	-0.0002	-0.0016	-0.0018
PSM	-0.0063	+0.0289	+0.0050
SWAT	+0.0050	+0.0052	+0.0052
GECCO	+0.0067	+0.0177	+0.0102
Macro	+0.0036	+0.0091	+0.0026

Table 19: Family-level VUS-PR contrasts relative to FITTA for the point-localization oracle variants.

Family	Witness	Normal mask	Both
GHL	-0.0008	-0.0165	-0.0160
CATSv2	+0.0049	-0.0039	-0.0019
PUMP	+0.0164	+0.0279	+0.0125
SMD	-0.0001	-0.0019	-0.0018
PSM	-0.0043	+0.0252	+0.0049
SWAT	+0.0064	-0.0021	-0.0004
GECCO	+0.0108	+0.0212	+0.0139
Macro	+0.0048	+0.0071	+0.0016

Oracle utilization. The point masks create adaptation opportunities that are unavailable to standard FITTA. Among 6,480 positive intervals, 2,319 contain at least one normal timestamp. Across the 79,530 detection units contained in positive intervals, 28,530 contain normal timestamps and therefore receive an additional masked-normal update in the normal-mask and combined oracle arms. These masks expose 327,921 normal timestamps for additional adaptation.

Table 20: Pooled utilization diagnostics for the point-localization oracle analysis. Rates in this table are pooled across entity–seed deployments rather than equal-family macro averages.

Quantity	Count or rate
Positive intervals	6,480
Positive intervals containing normal timestamps	2,319 (35.787%)
Positive detection units containing normal timestamps	28,530 / 79,530 (35.873%)
Additional masked-normal steps, normal mask / both	28,530 / 28,530
All-anomaly units skipped, normal mask / both	51,000 / 51,000
Normal timestamps unlocked	327,921
FITTA source-witness anomaly hit	5,807 / 6,480 (89.614%)
FITTA adapted-witness anomaly hit	5,831 / 6,480 (89.985%)
Source witness differs from oracle anomaly witness	673 / 6,480 (10.386%)
Adapted witness differs from oracle anomaly witness	649 / 6,480 (10.015%)

The source witness already falls on a true anomaly timestamp in most positive intervals, while normal masking unlocks a substantial number of additional confirmed-normal updates. Consistent with this utilization pattern, the normal-mask-only arm shows the largest aggregate improvement, whereas replacing the witness alone produces a smaller gain. The combined arm remaining close to FITTA shows that these two uses of additional information do not combine monotonically under the fixed objectives.

Scope of the diagnostic. This analysis uses simulated exact point labels revealed only after completed positive intervals. Point-level feedback contains more information than FITTA's existential bit, so matched response counts represent matched feedback timing rather than equal annotation effort or information cost. The analysis is restricted to the primary AE, the seven-family evaluation support, and the prescribed FITTA update objectives, with no oracle-specific hyperparameter tuning. It evaluates two specific uses of point labels and should therefore be interpreted as a localization-headroom diagnostic under the tested protocol, not as a performance ceiling for methods designed directly around point-level supervision.