Supplementary Material

This document provides supplementary material for the manuscript titled "*Rethinking the Unsupervised Paradigm in Time Series Anomaly Detection: An Empirical Study and Minimalist Baseline*".

A. Experimental Methodology and Benchmark Results Across Unbroken Sequences

To guarantee rigorous benchmarking and complete transparency, we outline the exact training and evaluation protocols across all baseline categories. Traditional unsupervised approaches (UTAD-I), including Random Guess, Isolation Forest, PCA, HBOS, and KMeans, operate in a strictly unsupervised manner by generating anomaly scores directly on the evaluation sequence without parameter optimization or label supervision, adhering to the standardized implementations in the TSB-UAD benchmark [1]. Deep reconstruction and normal-state modeling architectures (UTAD-II), covering AE, CNN, LSTM, TranAD, USAD, Omni, Anomaly Transformer, M2N2, and CATCH, optimize network weights exclusively on the uncontaminated normal training stream to characterize nominal operational patterns; during testing, anomaly scores are derived from prediction or reconstruction errors on the evaluation horizon. Finally, supervised classification baselines (STAD), comprising Random Forest, SVM, AdaBoost, ExtraTrees, LightGBM, and STAND, utilize a preceding labeled temporal window to learn discriminative decision boundaries directly from verified anomaly occurrences.

In the primary benchmark reported in Table 3 of the main manuscript, all models are evaluated on the identical, unobserved held-out horizon. At no point during training is the held-out evaluation stream exposed to any model. This design eliminates temporal data leakage and strictly tests real-world out-of-sample extrapolation.

For researchers seeking direct continuity with historical TSAD literature that evaluated models continuously across the unbroken test sequence, Table S1 reports the complete performance profiles across all five datasets and six metrics. Under this continuous evaluation, STAND and classical STAD ensembles maintain strong performance relative to unsupervised baselines, demonstrating that supervisory signals provide consistent benefits across both unbroken monitoring and out-of-sample temporal extrapolation.

The five evaluated multi-channel streams (PSM, SWaT, WADI, Swan, and Water) follow the standard benchmark definitions cataloged in TSB-UAD [1]. Baseline methods retain their official, benchmark-standardized architectures and hyperparameter configurations as implemented in TSB-UAD. To ensure exact reproducibility, the preprocessed continuous sequences and evaluation scripts are publicly archived in our benchmark repository (`https://github.com/EmorZz1G/SimAD`). Each telemetry channel is normalized via standard z-score normalization (`StandardScaler`):

$$\tilde{x}_{t,c} = \frac{x_{t,c} - \mu_c}{\sigma_c + \epsilon}, \quad c \in \{1, \ldots, C\} \tag{1}$$

where the channel-wise mean μ_c and standard deviation σ_c are fitted strictly on the training partition (the initial 50% prefix for STAD or nominal sequence for UTAD-II) and applied to the held-out evaluation suffix, eliminating look-ahead distribution leakage. Rare missing values are imputed via forward-fill followed by backward-fill.

B. Multi-Run Uncertainty, Seeds, and Effect Size Analysis

To evaluate statistical reproducibility and stability across stochastic initializations, all stochastic models and deep neural architectures were evaluated across $N = 5$ independent experimental runs initialized with distinct random seeds $\{40, 41, 42, 43, 44\}$. Table S3 presents the empirical standard deviations across all five benchmark datasets and six metrics evaluated on the held-out test horizon. Deterministic baselines including Isolation Forest, PCA, LOF, and HBOS exhibit zero variance across runs on fixed datasets. In contrast, deep neural architectures and stochastic tree ensembles exhibit moderate run-to-run variation. Across all evaluated benchmarks, the performance margin of supervised baselines over unsupervised alternatives substantially exceeds the observed standard deviations, confirming that the supervisory gain is statistically dependable.

Table S1
Performance comparison of the three method categories across five datasets and six metrics evaluated across the unbroken test sequence. **Bold** indicates the best performance, and underlining indicates the second-best performance. Methods highlighted in gray are UTAD-I, methods highlighted in green are UTAD-II, and methods highlighted in orange represent STAD methods.

Dataset	PSM						SWaT						WADI					
Metric	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR
Random	-0.83	8.51	67.99	5.53	50.10	32.99	-0.48	7.07	68.96	-3.13	49.96	12.77	-0.07	4.90	69.16	-6.51	49.80	8.47
IForest	-5.40	3.75	49.60	-1.69	46.76	29.96	-2.02	6.02	62.11	-9.86	32.74	10.37	7.83	10.99	61.99	-11.06	74.70	17.17
PCA	6.56	18.77	57.93	52.38	74.13	55.01	4.52	48.84	66.17	30.80	89.73	61.73	14.39	39.07	81.74	58.92	81.07	37.04
HBOS	-3.44	8.52	44.79	-18.88	49.90	33.23	9.65	10.38	71.49	24.54	75.08	26.34	9.44	9.86	67.40	6.38	73.97	18.50
KMeans	0.12	7.85	60.05	44.55	38.64	30.14	4.79	19.49	78.12	46.24	33.06	18.18	6.46	12.72	66.56	28.87	64.70	13.79
AE	7.78	5.19	56.50	51.07	56.50	34.90	16.25	17.87	74.95	40.34	84.69	45.08	15.04	27.00	74.26	42.28	64.38	16.65
CNN	0.57	26.33	69.16	60.48	61.26	52.52	29.04	**58.34**	5.46	5.46	86.77	60.84	2.39	24.20	78.57	43.04	66.08	24.00
LSTM	0.60	26.86	66.48	57.44	63.88	51.85	29.54	58.32	5.41	5.41	86.63	68.63	2.47	23.98	78.70	43.07	70.80	25.67
TranAD	0.23	20.09	63.50	58.40	60.54	46.85	29.78	**58.34**	5.34	5.34	88.47	63.50	2.60	23.23	77.80	40.01	72.76	30.27
USAD	1.49	21.39	46.14	44.90	61.53	46.62	29.79	**58.34**	5.33	5.33	88.71	63.04	8.36	23.01	77.83	39.86	73.60	28.02
Omni	3.93	21.78	23.44	23.04	64.11	44.55	29.80	**58.34**	5.31	5.31	88.85	59.30	10.84	39.02	72.07	52.59	79.13	38.82
A.T.	1.02	18.20	48.73	42.56	59.13	43.46	29.31	**58.34**	5.35	5.35	74.39	47.13	3.88	28.02	74.06	22.00	61.97	20.27
M2N2	-0.07	15.26	79.11	47.11	63.35	48.17	0.01	23.99	75.65	42.15	81.15	33.29	1.69	20.75	74.89	22.97	64.38	22.25
CATCH	0.62	16.85	**84.85**	62.92	60.65	51.99	0.02	9.33	73.41	23.84	29.78	11.87	0.26	13.64	74.92	31.67	69.76	21.18
RF	44.33	64.86	76.49	69.43	90.77	80.86	61.42	56.73	86.06	85.08	95.92	89.24	61.83	**87.71**	90.38	88.91	97.58	91.77
SVM	15.66	29.80	22.34	21.81	91.93	80.46	26.95	35.33	81.49	80.25	93.29	88.62	32.21	80.41	**92.49**	84.41	93.89	87.95
AdaBoost	14.46	30.48	29.28	28.95	84.99	75.80	18.69	52.01	50.13	42.59	86.71	41.43	24.53	78.40	83.81	78.44	91.84	75.43
ExtraTrees	**50.28**	**66.16**	78.92	**73.20**	**92.82**	**83.89**	69.03	53.87	81.26	81.22	**97.01**	**92.69**	63.07	84.94	87.16	79.12	**99.10**	**95.48**
LightGBM	44.21	30.53	12.18	12.18	87.36	77.24	66.51	48.82	75.89	70.81	94.87	80.76	**66.43**	80.95	91.93	**89.76**	96.89	92.03
STAND	46.94	30.33	29.65	29.35	90.74	83.66	**71.48**	57.95	**87.84**	**87.42**	92.02	87.14	65.24	82.03	89.93	84.07	91.72	87.08
Dataset	Swan						Water						Average					
Metric	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR
Random	-0.27	8.65	26.68	-0.69	49.79	83.90	0.72	1.53	63.75	-3.53	50.32	3.57	-0.18	6.13	59.31	-1.67	49.99	28.34
IForest	5.85	18.18	11.56	5.15	66.30	85.11	12.35	8.11	60.59	19.21	79.10	10.09	3.72	9.41	49.17	0.35	59.92	30.54
PCA	1.55	24.75	2.21	2.00	59.62	93.45	11.38	18.03	62.58	47.28	90.68	17.92	7.68	29.89	54.12	38.28	79.05	53.03
HBOS	15.92	18.01	23.08	20.14	78.82	88.77	16.45	8.16	65.15	9.51	82.93	11.87	9.60	10.99	54.38	8.34	72.14	35.74
KMeans	-7.41	6.43	11.18	6.27	30.04	81.24	28.38	17.89	72.45	46.33	92.85	24.01	6.47	12.88	57.67	34.45	51.86	33.47
AE	5.95	18.91	21.89	20.80	60.01	88.57	1.76	5.68	54.44	22.81	50.71	2.36	9.36	14.93	56.41	35.46	63.26	37.51
CNN	0.56	24.84	15.92	15.78	72.99	93.71	2.62	3.96	44.78	9.19	54.04	2.55	7.04	27.53	42.78	26.79	68.23	46.72
LSTM	0.38	23.98	27.41	27.11	75.24	92.63	2.62	8.63	54.09	12.65	60.35	5.84	7.12	28.36	46.42	29.14	71.38	48.92
TranAD	0.08	15.94	18.57	5.76	65.01	91.96	0.32	6.20	50.82	13.96	52.11	6.51	6.60	24.76	43.21	24.69	67.78	47.82
USAD	0.54	20.85	11.95	11.42	61.81	92.27	3.21	9.35	41.21	13.84	71.66	6.27	8.68	26.59	36.49	23.07	71.46	47.25
Omni	1.86	24.44	0.26	-0.25	68.55	94.25	3.23	9.30	41.19	5.74	74.37	6.44	9.93	30.58	28.45	17.29	75.00	48.67
A.T.	0.02	3.45	6.64	-5.31	50.27	84.22	1.91	3.29	71.44	30.10	65.94	5.02	7.23	22.26	41.24	18.94	62.34	40.02
M2N2	0.42	26.50	0.03	0.03	77.54	96.62	0.65	22.13	84.87	63.64	91.91	37.77	0.54	21.73	62.91	35.18	75.67	47.62
CATCH	0.22	25.60	3.05	3.03	61.55	94.17	0.88	18.98	81.54	54.51	91.28	27.36	0.40	16.88	63.55	35.19	62.61	41.31
RF	62.27	**45.36**	**81.28**	**80.84**	**96.39**	**97.94**	70.84	39.31	81.52	78.35	93.55	55.41	50.75	37.69	**71.85**	66.12	**92.29**	75.20
SVM	21.99	25.79	51.55	51.29	89.07	97.04	51.56	32.48	**98.42**	**96.81**	98.55	82.80	21.22	28.20	61.82	53.88	86.47	67.16
AdaBoost	27.03	25.56	41.22	41.03	80.13	92.86	47.29	32.58	92.36	82.25	98.83	88.42	23.21	29.25	53.55	44.76	82.25	61.13
ExtraTrees	57.48	41.98	74.27	73.10	95.58	97.51	**92.52**	35.80	98.42	96.62	**99.76**	93.93	50.19	34.57	69.58	60.87	91.47	73.90
LightGBM	**67.56**	25.14	46.38	46.04	94.28	97.34	71.07	**73.33**	82.77	79.06	94.11	54.01	47.94	32.42	63.68	51.93	88.72	68.99
STAND	55.53	25.57	46.98	46.65	84.76	97.19	84.44	34.01	96.61	93.41	99.62	**95.38**	**64.73**	**45.98**	70.20	**68.18**	91.77	**90.09**

B.1. Explicit Effect Size Analysis (Cohen's d)

To provide an explicit statistical effect size assessment beyond empirical standard deviations, we compute the standardized mean difference using Cohen's d between STAND and representative unsupervised deep learning baselines across the five evaluated benchmarks:

$$d = \frac{\bar{X}_{\text{STAND}} - \bar{X}_{\text{baseline}}}{s_{\text{pooled}}}, \quad \text{where} \quad s_{\text{pooled}} = \sqrt{\frac{s^2_{\text{STAND}} + s^2_{\text{baseline}}}{2}}. \tag{2}$$

Standard statistical conventions characterize $|d| \geq 0.2$ as a small effect, $|d| \geq 0.5$ as a medium effect, $|d| \geq 0.8$ as a large effect, and $|d| \geq 1.2$ as a very large effect. Table S2 summarizes the resulting Cohen's d values.

On prediction certainty (CCE), STAND establishes substantial and decisive positive effect sizes against all deep unsupervised baselines, ranging from $d = +1.40$ (vs. USAD) to $d = +2.77$ (vs. Anomaly Transformer), demonstrating that the observed prediction confidence margin is statistically pronounced relative to operational variance. On global discriminative ranking (AUC-ROC), STAND yields large positive effect sizes against CATCH ($d = +1.03$), Anomaly Transformer ($d = +3.40$), and AE ($d = +0.99$). Even on threshold-dependent point-wise classification (Best-F1), STAND exhibits substantial positive effect sizes against CATCH ($d = +1.14$), Anomaly Transformer ($d = +1.45$), and AE ($d = +0.67$), confirming that supervisory signals deliver statistically dependable empirical advantages across benchmarks.

Table S2
Standardized effect sizes (Cohen's d) measuring the empirical advantage of STAND relative to representative unsupervised deep learning baselines across benchmark metrics. Values $|d| \geq 0.8$ denote large statistical effect sizes.

Comparator	CCE	F1	Aff-F1	UAff-F1	AUC-ROC	VUS-PR
vs. CATCH	+2.51	+1.14	-0.11	+0.93	+1.03	+0.45
vs. Anomaly Transformer	+2.77	+1.45	+0.03	+1.55	+3.40	+0.86
vs. OmniAnomaly	+1.55	+0.10	+0.51	+0.49	+0.33	-0.05
vs. TranAD	+1.61	+0.24	+0.21	+0.28	+0.29	+0.11
vs. USAD	+1.40	+0.11	+0.33	+0.26	+0.23	+0.11
vs. LSTM	+1.42	+0.01	+0.08	+0.13	+0.03	+0.09
vs. AutoEncoder (AE)	+1.90	+0.67	+0.09	+0.90	+0.99	+0.43

C. Non-Parametric Statistical Ranking Analysis Across Benchmarks

To complement the arithmetic score averages reported in the main manuscript with a scale-invariant assessment, this section presents a comprehensive non-parametric ranking analysis across all 30 evaluation scenarios covering five datasets and six metrics. For each dataset and metric combination, all 20 evaluated models are assigned ordinal ranks from 1 to 20, with ties resolved via standard fractional mid-ranks.

Table S4 reports the metric-specific average ranks across datasets, alongside the overall average rank and final placement. STAND achieves the top overall average rank of 5.70, outperforming strong tree ensembles such as ExtraTrees at 5.87 and Random Forest at 6.93, as well as the best-performing unsupervised baselines such as CNN at 7.77, LSTM at 7.97, and PCA at 8.17. At the paradigm level, supervised methods attain an average rank of 7.60, substantially outperforming deep normal-state modeling architectures at 11.04 and classical unsupervised detectors at 13.02. These ranking results demonstrate that the empirical advantages of supervisory guidance and STAND remain robust across disparate metric scales without being influenced by outlier values.

D. Full Results of STAND Sensitivity Analysis

To investigate the parameter sensitivity of STAND, we conduct a comprehensive parameter analysis across all five real-world datasets and the three PSM subsets. Rather than freezing remaining hyperparameters to arbitrary baseline values, each hyperparameter level is evaluated via marginal sensitivity analysis across the configuration grid (varying temporal window sizes $W \in \{32, 64, 128\}$ and number of TEM layers $L \in \{1, 2, 3\}$) across independent runs with distinct random initializations, where bars and error bars denote the marginal sample mean and 95% confidence intervals.

Figures S1, S2, and S3 illustrate the impact of the model dimension d_{model}, the number of TEM layers, and the temporal window size on STAND's performance across the five datasets, respectively. Overall, performance improves across nearly all metrics as the model dimension increases. Conversely, performance tends to degrade across point-wise and event-level metrics as the number of TEM layers and the temporal window size increase.

Figures S4, S5, and S6 present the impact of d_{model}, the number of TEM layers, and the window size on STAND's performance across the three PSM subsets. Generally, the performance trend is consistent with that observed across the five real-world datasets. Specifically, when the model dimension increases, performance improves across subsets SP2 and SP3, while the performance variation on SP1 remains comparatively marginal. Conversely, when the number of TEM layers and the temporal window size increase, performance slightly decreases on the SP2 and SP3 subsets, whereas no pronounced downward trend emerges on SP1.

Therefore, the influence of hyperparameters on STAND is primarily modulated by training data volume. When larger volumes of supervisory data are available, models exhibit stable and coherent parametric trajectories, whereas under severely restricted label regimes, stochastic initialization induces greater dispersion.

Table S3
Multi-run standard deviations of the evaluated methods across five datasets and six metrics on the held-out test horizon. Standard deviations for individual datasets are computed across independent random runs, while the Average column reports the cross-dataset standard deviation. Methods highlighted in gray are UTAD-I, methods highlighted in green are UTAD-II, and methods highlighted in orange represent STAD methods.

Dataset	PSM						SWaT						WADI					
Metric	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR
Random	0.54	0.28	0.20	3.54	0.13	0.20	0.25	0.14	0.64	0.23	0.05	0.04	0.33	0.60	0.85	28.15	1.14	0.20
IForest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LOF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PCA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HBOS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KMeans	0.45	0.30	0.98	1.23	0.96	0.32	0.70	1.30	0.63	1.57	0.40	0.91	2.64	0.68	5.78	26.14	3.50	1.15
AE	0.04	0.01	0.00	3.71	0.09	0.05	0.36	0.37	0.45	1.07	0.74	0.93	0.09	0.00	0.00	1.03	0.70	0.32
CNN	0.03	1.02	2.24	6.25	0.13	1.33	0.10	0.09	0.45	0.45	0.73	0.50	0.08	0.38	0.05	0.88	0.32	1.10
LSTM	0.01	5.79	0.88	0.69	0.71	3.92	0.01	0.01	0.11	0.11	0.17	0.59	0.32	0.08	0.08	7.84	2.14	2.20
TranAD	0.02	0.18	0.72	2.52	0.44	0.59	0.02	0.01	0.16	0.16	0.07	0.02	0.35	0.08	0.03	0.58	1.26	1.09
USAD	0.06	0.46	0.07	0.27	0.56	0.46	0.01	0.00	0.02	0.02	0.10	0.23	0.23	0.08	0.09	6.07	3.84	1.23
Omni	0.02	0.02	0.00	0.17	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.01	2.61	0.04	0.05
A.T.	0.06	1.53	1.32	7.84	0.80	0.65	0.00	3.77	2.41	18.36	17.50	1.20	3.47	9.50	5.00	18.66	5.05	5.20
M2N2	0.00	0.93	0.23	2.69	0.62	1.35	0.09	0.99	0.08	0.24	0.48	0.32	0.10	0.13	0.12	2.87	1.53	1.30
CATCH	0.31	0.37	0.59	2.16	0.05	0.23	0.00	0.14	0.14	3.49	0.23	0.18	0.27	1.03	0.62	0.84	0.13	1.15
RF	2.21	4.23	9.13	11.84	4.04	2.88	1.98	9.23	12.43	11.21	1.49	14.40	0.52	3.83	7.89	19.19	3.68	6.46
SVM	2.32	2.48	3.04	4.81	3.31	0.02	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	10.27	0.00	0.00
AdaBoost	0.69	1.63	16.83	7.35	0.87	7.79	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.58	0.52	2.69	0.62	3.66
ExtraTrees	0.58	2.12	6.14	5.34	3.35	1.30	4.79	16.93	4.21	6.22	1.78	5.09	2.13	5.10	8.10	10.16	1.07	4.56
LightGBM	1.35	2.61	3.80	2.12	2.91	3.22	0.98	13.53	10.35	21.63	1.11	4.26	1.72	3.38	2.14	8.82	4.26	1.55
STAND	1.61	1.53	5.07	1.86	0.26	0.04	1.30	2.52	14.29	10.94	2.37	3.68	0.84	5.96	7.92	10.39	7.96	0.87

Dataset	Swan						Water						Average					
Metric	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR	CCE	F1	Aff-F1	UAff-F1	AUC	V-PR
Random	0.30	0.17	0.33	7.00	0.17	0.45	3.45	0.28	0.67	5.42	3.81	0.10	1.77	3.74	18.53	12.21	2.16	35.85
IForest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.27	4.37	27.40	26.44	26.53	32.51
LOF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.75	3.82	22.30	12.03	15.40	35.78
PCA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.94	11.35	32.61	28.82	17.57	27.56
HBOS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.45	5.60	22.63	28.93	18.10	33.00
KMeans	0.32	0.50	0.44	1.74	1.51	0.33	1.30	0.06	2.53	4.88	0.44	0.42	22.22	3.44	25.41	46.37	33.68	30.39
AE	0.10	0.04	0.09	0.08	0.19	0.01	0.08	0.00	0.00	2.13	0.05	0.00	6.55	12.56	25.52	15.58	14.07	37.87
CNN	0.01	0.03	0.00	0.00	1.39	0.03	1.11	0.38	0.86	2.11	0.50	7.34	16.87	12.20	41.27	34.20	12.42	22.78
LSTM	0.00	0.02	0.17	0.17	0.67	0.17	0.60	1.15	0.69	1.75	3.35	1.91	16.64	13.62	34.96	27.39	9.27	31.62
TranAD	0.00	0.04	0.09	0.68	0.58	0.06	0.01	0.13	0.29	0.56	0.51	0.19	15.80	14.50	36.23	37.25	14.90	30.43
USAD	0.00	0.00	0.03	0.05	0.83	0.06	0.02	0.00	0.11	0.70	0.21	0.03	15.97	13.93	40.39	32.38	15.75	31.18
Omni	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.21	0.02	0.02	13.25	14.18	42.43	34.26	14.18	23.63
A.T.	0.06	5.27	6.64	10.65	3.22	0.75	9.21	1.61	14.58	47.90	7.99	1.37	6.91	6.62	22.08	35.18	8.46	34.96
M2N2	0.00	0.02	0.00	0.00	0.12	0.04	0.53	0.22	0.54	2.04	0.71	0.46	5.32	7.60	36.51	22.58	9.38	28.88
CATCH	0.02	0.00	0.00	0.00	0.25	0.01	2.88	0.11	0.41	3.31	1.39	1.36	10.16	6.04	34.29	21.53	27.63	31.31
RF	0.58	1.33	1.20	2.38	0.17	0.80	3.44	6.91	11.39	3.42	6.44	4.33	17.49	13.93	24.95	19.68	9.20	27.40
SVM	1.15	1.76	2.37	1.23	2.41	1.92	0.58	0.43	5.84	14.38	2.22	4.37	7.51	10.68	26.25	27.29	6.56	33.98
AdaBoost	2.11	5.13	8.02	8.04	0.18	1.12	0.00	0.00	0.00	0.74	0.00	0.00	14.99	10.98	27.63	20.93	10.02	24.86
ExtraTrees	0.26	0.71	8.82	8.41	0.57	0.17	2.70	0.25	1.94	5.09	0.56	9.55	20.53	18.61	26.17	19.71	9.81	18.08
LightGBM	1.04	0.17	5.35	5.33	0.40	0.16	0.03	0.59	5.58	23.71	0.88	19.43	27.22	16.44	19.92	18.78	12.27	22.25
STAND	1.55	0.57	12.43	12.23	0.77	0.40	1.40	1.03	1.13	4.35	1.14	7.44	17.20	13.19	24.91	22.96	10.67	31.42

Additionally, average Aff-F1 scores remain higher on SP1 than on SP2 and SP3. This occurs because buffer-based affinity metrics grant tolerance to predictions adjacent to anomaly boundaries, which tends to yield elevated scores when annotated anomalies are scarce.

E. Qualitative Analysis on the Training Prefix

To complement the out-of-sample visual analysis presented in the main text, Figure S7 displays the anomaly score trajectories across an illustrative segment of the training prefix from timestamp 150k to 200k on SWaT.

During this training segment, two distinct anomaly intervals occur near timestamp 173k and timestamp 199k. Supervised models, including ExtraTrees, LightGBM, and STAND, reliably produce scores close to one during both intervals while maintaining a stable zero baseline across normal periods. In contrast, the unsupervised models LSTM and M2N2 output near-zero anomaly scores throughout this entire window, failing to register either event despite having access to the unlabeled operational data. This observation confirms that unsupervised reconstruction errors often struggle to separate subtle anomalous deviations from complex multi-channel background dynamics, whereas direct supervisory guidance provides clear discriminative boundaries.

Table S4
Non-parametric ranking summary of all 20 models across 30 evaluation settings covering five datasets and six metrics. Ranks range from 1 (best) to 20 (worst), with ties resolved using fractional mid-ranks. Metric columns report the average rank across the five datasets for each metric. Overall Avg. Rk. denotes the mean rank across all 30 evaluation scenarios, and Position indicates the overall placement. Bold font indicates the best performance, and underlining indicates the second-best performance. Shaded rows highlight UTAD-I (gray), UTAD-II (green), and STAD (orange) methods.

Model	CCE	F1	Aff-F1	UAff-F1	AUC-ROC	VUS-PR	Overall Avg. Rk.	Position
Random	19.20	18.20	9.20	15.80	18.20	18.40	16.50	19
IForest	12.40	12.60	14.00	17.80	10.40	16.20	13.90	17
PCA	8.00	7.40	9.80	10.40	6.80	6.60	8.17	6
HBOS	10.40	14.80	11.00	13.00	10.00	13.20	12.07	14
KMeans	14.20	16.60	11.40	11.20	16.80	16.60	14.47	18
AE	9.40	15.80	11.00	10.40	13.20	15.40	12.53	16
CNN	10.20	5.00	12.50	10.70	**4.60**	**3.60**	7.77	4
LSTM	10.80	8.40	7.20	5.60	8.80	7.00	7.97	5
TranAD	13.20	11.80	10.80	9.20	8.00	8.40	10.23	12
USAD	9.00	9.30	12.00	8.20	9.80	8.60	9.48	9
Omni	8.60	8.40	14.30	12.10	9.00	6.80	9.87	11
A.T.	17.60	18.80	13.20	19.60	18.20	17.80	17.53	20
M2N2	17.40	10.90	10.70	11.50	10.60	8.40	11.58	13
CATCH	15.20	12.50	9.10	11.30	14.80	11.20	12.35	15
RF	4.20	5.60	11.20	6.60	6.80	7.20	6.93	3
SVM	9.80	8.40	8.20	7.40	8.40	7.00	8.20	7
AdaBoost	8.60	6.90	11.40	8.00	10.80	12.40	9.68	10
ExtraTrees	4.80	**3.60**	6.80	5.20	6.60	8.20	5.87	2
LightGBM	5.60	8.20	10.80	12.00	8.00	10.60	9.20	8
STAND	**1.40**	6.80	**5.40**	**4.00**	10.20	6.40	**5.70**	**1**
UTAD-I Average	12.84	13.92	11.08	13.64	12.44	14.20	13.02	–
UTAD-II Average	12.38	11.21	11.20	10.96	10.78	9.69	11.04	–
STAD Average	5.73	6.58	8.97	7.20	8.47	8.63	7.60	–

Table S5
Aff-F1 performance under varying positive class weights w_{pos} in BCE loss on the Water dataset (1.05% anomaly ratio). Results report mean $\pm$ standard deviation across multiple repeated runs.

Positive Weight w_{pos}	**1.0 (Default)**	**4.0**	**16.0**	**64.0**
Aff-F1	92.21 ± 1.14	90.78 ± 0.76	90.80 ± 0.63	91.38 ± 1.64

F. Ablation on Class Imbalance Weighting

Industrial multivariate time series typically exhibit severe class imbalance; for instance, in the Water benchmark, anomalies account for only 1.05% of all timestamps. In our default setting, STAND is trained using the standard unweighted Binary Cross-Entropy (BCE) loss. To empirically evaluate whether positive class re-weighting benefits detection under sparse anomalies, we consider the weighted BCE loss:

$$\mathcal{L}_w(\mathbf{S}, \mathbf{y}) = -\frac{1}{T}\sum_{t=1}^{T}\left[w_{\text{pos}} \cdot y_t \log(\sigma(s_t)) + (1 - y_t)\log(1 - \sigma(s_t))\right], \tag{3}$$

where $w_{\text{pos}} \geq 1$ explicitly scales the loss penalty on anomalous timestamps ($y_t = 1$, corresponding to the `pos_weight` parameter in PyTorch's `BCEWithLogitsLoss`), and $w_{\text{pos}} = 1.0$ recovers our default unweighted objective. We conduct an ablation by varying $w_{\text{pos}} \in \{1, 4, 16, 64\}$. Table S5 reports the resulting Aff-F1 performance across multiple repeated runs (mean $\pm$ standard deviation) on the Water dataset.

Across this 64-fold weighting range, detection performance remains relatively stable within a narrow interval of $90.78 \sim 92.21$, demonstrating that STAND's sequence-level temporal representations are reasonably robust to class weighting. Furthermore, increasing the positive penalty does not yield a consistent or monotonic performance advantage, with the unweighted baseline ($w_{\text{pos}} = 1.0$) attaining a slightly higher score (92.21). Because aggressive loss weighting introduces hyperparameter sensitivity and risks inflating

boundary false alarms on industrial telemetry without demonstrable empirical gains, retaining the standard unweighted BCE loss adheres to the principle of parsimony while preserving baseline simplicity.

G. Extended Label Noise Robustness Profiles with Error Bands

To examine the resilience of supervised anomaly detection models against contiguous interval misclassifications and systematic label corruptions, Figure S8 reports the complete performance profiles under varying ratios of contiguous Segment Noise ($\rho \in [0.0, 0.10]$) on the PSM benchmark.

In this evaluation, contiguous blocks of normal timestamps are systematically inverted to anomalous until reaching the specified corruption ratio ρ, simulating operational cycle misclassifications, delayed human annotator reaction, and boundary ambiguities. To ensure statistical reliability and prevent findings from depending on an isolated perturbation pattern, each noise setting is evaluated across multiple repeated runs with distinct random corruption realizations. The solid curves plot the empirical mean, while shaded error bands denote the standard deviation ($\pm 1\sigma$) across multiple independent realizations.

Several key observations emerge from the empirical results:

1. **Robustness on CCE and F1**: On CCE, STAND consistently records the highest scores across all tested corruption levels ($7.7 \sim 16.4$), remaining substantially higher than the unsupervised CATCH baseline (0.54) even at the highest corruption ratio ($\rho = 0.10$). Similarly, on F1, STAND remains competitive across the noise spectrum, maintaining an average score between 15.1 and 23.6 in comparison to CATCH (15.26).
2. **Crossover dynamics on UAff-F1**: Contiguous segment corruptions alter local boundary statistics more severely than isolated point noise. For UAff-F1, STAND's performance experiences a downward trajectory as corrupted segments expand, crossing below the unsupervised CATCH baseline (56.61) around $\rho \approx 0.03 \sim 0.05$. This empirical crossover characterizes the regime where unsupervised reconstruction modeling becomes advantageous for localized event alignment.
3. **Global discriminative ranking**: On AUC-ROC and VUS-PR, supervised models (particularly ExtraTrees, SVM, and STAND) maintain stable ranking capability, with ExtraTrees and SVM comfortably exceeding the unsupervised threshold across all evaluated segment noise levels up to $\rho = 0.10$.

H. Discussion on Limited-Supervision and Imperfect-Label Paradigms

To complement the literature positioning in Section 2.1, this section provides an extended methodological discussion distinguishing STAND from related limited-supervision and imperfect-label paradigms.

Semi-supervised and positive-unlabeled approaches address label scarcity through unverified data assumptions. Semi-supervised methods, such as Deep SAD [3], DevNet [4], and recent outlier-fraction analyses [5], commonly assume transductive access to the target sequence distribution during training or restrict training strictly to nominal sequences (recovering one-class UTAD-II). Positive-unlabeled (PU) learning [6] operates when only anomalous instances are verified, treating all remaining data as an unlabeled mixture and requiring sensitive estimation of the latent anomaly prior $\pi = P(y = 1)$. In contrast, STAND operates under a strict inductive temporal protocol: training utilizes a localized historical prefix with confirmed binary labels, while the subsequent evaluation horizon remains entirely unobserved during training. This eliminates transductive distribution leakage and avoids fragile class-prior estimation under severe class imbalance.

Weakly-supervised, active, and few-shot formulations introduce specialized algorithmic or operational workflows to compensate for scarce labels. Weakly-supervised methods, such as multiple-instance learning (MIL) frameworks [7], address label coarseness where only sequence-level bag tags are available, requiring specialized attention pooling or ranking losses to infer temporal boundaries. Active learning architectures [8] incorporate interactive human-in-the-loop querying during inference, introducing continuous query latency and human intervention that are impractical for unattended industrial automation. Few-shot and meta-learning paradigms [9] rely on pretraining transferable metric spaces over diverse pools of auxiliary source datasets, which are rarely available across unaligned industrial telemetry systems. In contrast, STAND addresses label sparsity across time rather than label imprecision: within the observed historical prefix,

Table S6
Impact of event-level few-shot sample budgets (k) on model performance across CCE, F1, and VUS-PR metrics on PSM and SWaT.

Dataset	Metric	Model	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$	$k=11$	$k=13$	$k=15$
PSM	CCE	**STAND (Ours)**	-0.80	-0.97	1.03	8.03	9.81	11.17	15.36	11.42
PSM	CCE	ExtraTrees	5.55	4.73	9.14	2.92	10.20	8.00	8.82	9.81
PSM	CCE	SVM	0.93	-0.59	5.75	4.67	9.11	9.46	8.75	9.14
PSM	CCE	AdaBoost	5.34	3.51	4.47	3.32	4.16	8.75	6.90	7.25
PSM	F1	**STAND (Ours)**	9.04	8.84	17.11	22.53	20.69	22.39	19.48	21.25
PSM	F1	ExtraTrees	22.71	20.97	25.53	4.49	18.73	16.81	14.87	14.44
PSM	F1	SVM	2.11	4.77	21.78	3.46	16.42	21.61	18.35	18.53
PSM	F1	AdaBoost	23.01	15.22	21.74	10.32	21.36	24.21	17.95	19.60
PSM	VUS-PR	**STAND (Ours)**	33.50	33.17	43.27	44.22	49.07	53.78	52.57	50.44
PSM	VUS-PR	ExtraTrees	46.71	46.75	55.27	36.04	56.18	52.59	52.09	53.17
PSM	VUS-PR	SVM	35.32	32.61	55.65	39.66	54.45	57.21	54.58	54.92
PSM	VUS-PR	AdaBoost	53.47	43.46	43.33	38.56	43.86	53.62	47.71	49.46
SWaT	CCE	**STAND (Ours)**	-1.81	-0.57	3.84	16.70	31.65	26.52	30.64	24.57
SWaT	CCE	ExtraTrees	-4.41	9.55	4.56	6.56	44.71	42.83	42.90	39.24
SWaT	CCE	SVM	5.54	11.32	10.81	15.77	32.23	35.42	24.16	15.48
SWaT	CCE	AdaBoost	-4.92	-4.28	-5.95	-10.95	15.95	17.78	13.81	15.12
SWaT	F1	**STAND (Ours)**	3.62	2.37	4.38	9.03	37.68	5.79	40.76	34.42
SWaT	F1	ExtraTrees	5.54	3.87	2.31	6.07	29.39	86.87	86.40	86.17
SWaT	F1	SVM	0.81	7.18	8.15	18.68	40.83	40.92	40.21	32.99
SWaT	F1	AdaBoost	2.52	6.53	8.64	6.55	0.02	0.00	0.62	0.63
SWaT	VUS-PR	**STAND (Ours)**	13.30	29.83	18.42	35.88	51.92	34.59	48.65	43.15
SWaT	VUS-PR	ExtraTrees	16.52	33.61	21.94	20.78	67.80	62.69	65.01	70.66
SWaT	VUS-PR	SVM	30.09	50.59	42.30	56.59	83.23	82.91	79.97	66.05
SWaT	VUS-PR	AdaBoost	17.13	18.68	17.00	14.39	48.26	53.45	55.76	52.82

timestamp-level transitions are precisely annotated, allowing standard sequence discrimination trained from scratch on local telemetry without heuristic bag pooling, recurring human queries, or cross-task meta-training.

Regarding benchmark comparability, directly evaluating these diverse paradigms on an identical benchmark table involves incompatible operational assumptions, as each paradigm operates under fundamentally distinct input interfaces (such as coarse bag tags in MIL, unverified negative pools in PU learning, interactive query budgets in active learning, and cross-task source pools in few-shot learning). By evaluating STAND against 20 competitive unsupervised and semi-supervised baselines under a standardized inductive temporal split across five benchmark datasets, our empirical study maintains strict comparability while establishing a clear baseline for the utility of direct temporal supervision.

I. Detailed Event-Level Few-Shot Evaluation

To supplement the event-level few-shot analysis in Section 5.4.1, Table S6 provides the complete numerical performance measurements across all evaluated supervision budgets $k \in \{1, 3, 5, 7, 9, 11, 13, 15\}$ on the PSM and SWaT benchmarks. For each configuration, training partitions contain exactly k contiguous failure events alongside preceding nominal operational context.

J. Comprehensive Matched-Backbone Comparison Across Benchmarks

To provide complete empirical transparency for Section 5.6.3, Table S7 and Table S8 present the detailed benchmark-by-benchmark evaluation across all five datasets (PSM, SWaT, WADI, Swan, Water) comparing identical sequence backbones (Bi-LSTM, Attention, Bi-GRU) under unsupervised reconstruction (UTAD,

Table S7
Detailed cross-paradigm evaluation comparing identical sequential architectures across PSM, SWaT, and WADI. UTAD denotes unsupervised reconstruction using MSE loss with 0 labels; STAD denotes supervised classification using BCE loss. Net supervisory gain ($\Delta = \text{STAD} - \text{UTAD}$) isolates the empirical return of the learning objective under matched sequential capacity.

Encoder	Objective	CCE	F1	Aff-F1	UAff-F1	AUC-ROC	VUS-PR
PSM							
Bi-LSTM	UTAD (MSE)	-0.08	12.74	74.53	25.07	55.66	41.80
	STAD (BCE)	16.39	23.58	64.18	55.61	66.81	55.47
	Gain (Δ)	*+16.47*	*+10.84*	*-10.35*	*+30.54*	*+11.15*	*+13.67*
Attention	UTAD (MSE)	-0.08	13.28	75.71	30.93	56.46	42.65
	STAD (BCE)	7.02	10.28	59.97	15.55	55.88	39.56
	Gain (Δ)	*+7.10*	*-3.00*	*-15.74*	*-15.38*	*-0.58*	*-3.09*
Bi-GRU	UTAD (MSE)	-0.08	12.64	75.21	31.41	54.79	41.07
	STAD (BCE)	11.58	15.63	61.55	35.79	62.70	46.98
	Gain (Δ)	*+11.66*	*+2.99*	*-13.66*	*+4.38*	*+7.91*	*+5.91*
SWaT							
Bi-LSTM	UTAD (MSE)	1.36	38.20	78.26	39.99	76.49	55.99
	STAD (BCE)	51.97	36.88	76.94	69.73	90.68	79.62
	Gain (Δ)	*+50.61*	*-1.32*	*-1.32*	*+29.74*	*+14.19*	*+23.63*
Attention	UTAD (MSE)	1.29	38.68	78.63	39.73	74.98	53.75
	STAD (BCE)	47.36	41.30	43.29	43.18	90.02	85.59
	Gain (Δ)	*+46.07*	*+2.62*	*-35.34*	*+3.45*	*+15.04*	*+31.84*
Bi-GRU	UTAD (MSE)	1.36	39.16	78.52	39.62	73.79	54.79
	STAD (BCE)	25.71	41.50	34.59	34.43	90.46	80.67
	Gain (Δ)	*+24.35*	*+2.34*	*-43.93*	*-5.19*	*+16.67*	*+25.88*
WADI							
Bi-LSTM	UTAD (MSE)	0.15	4.63	66.38	-18.03	59.10	8.68
	STAD (BCE)	26.70	37.33	78.10	55.51	73.17	39.71
	Gain (Δ)	*+26.55*	*+32.70*	*+11.72*	*+73.54*	*+14.07*	*+31.03*
Attention	UTAD (MSE)	0.12	4.85	66.90	-28.25	59.72	8.65
	STAD (BCE)	12.39	21.05	85.34	63.32	71.92	18.93
	Gain (Δ)	*+12.27*	*+16.20*	*+18.44*	*+91.57*	*+12.20*	*+10.28*
Bi-GRU	UTAD (MSE)	0.14	4.94	66.12	-5.06	58.90	8.69
	STAD (BCE)	16.93	18.66	79.44	62.86	81.06	19.38
	Gain (Δ)	*+16.79*	*+13.72*	*+13.32*	*+67.92*	*+22.16*	*+10.69*

MSE loss, 0 anomaly labels) and supervised classification (STAD, BCE loss, full training labels), averaged across independent experimental runs. Net supervisory gains ($\Delta = \text{STAD} - \text{UTAD}$) quantify the performance boost directly attributable to supervisory feedback.

Table S8
Detailed cross-paradigm evaluation comparing identical sequential architectures across Swan and Water. UTAD denotes unsupervised reconstruction using MSE loss with 0 labels; STAD denotes supervised classification using BCE loss. Net supervisory gain (Δ = STAD − UTAD) isolates the empirical return of the learning objective under matched sequential capacity.

Encoder	Objective	CCE	F1	Aff-F1	UAff-F1	AUC-ROC	VUS-PR
Swan							
Bi-LSTM	UTAD (MSE)	0.03	9.36	17.11	-0.82	40.39	92.00
	STAD (BCE)	33.26	17.17	27.39	26.95	77.27	96.69
	Gain (Δ)	*+33.23*	*+7.81*	*+10.28*	*+27.77*	*+36.88*	*+4.69*
Attention	UTAD (MSE)	0.03	8.59	18.06	-4.66	39.69	91.81
	STAD (BCE)	33.32	15.95	30.19	28.77	81.60	95.37
	Gain (Δ)	*+33.29*	*+7.36*	*+12.13*	*+33.43*	*+41.91*	*+3.56*
Bi-GRU	UTAD (MSE)	0.03	9.32	17.10	-0.45	39.93	91.90
	STAD (BCE)	31.16	16.06	25.81	24.70	80.94	95.46
	Gain (Δ)	*+31.13*	*+6.74*	*+8.71*	*+25.15*	*+41.01*	*+3.56*
Water							
Bi-LSTM	UTAD (MSE)	4.21	2.80	84.61	63.64	74.31	4.74
	STAD (BCE)	64.19	6.63	92.21	83.33	96.53	20.37
	Gain (Δ)	*+59.98*	*+3.83*	*+7.60*	*+19.69*	*+22.22*	*+15.63*
Attention	UTAD (MSE)	4.05	2.54	85.91	67.62	72.32	4.68
	STAD (BCE)	27.03	1.44	86.54	68.76	79.50	1.17
	Gain (Δ)	*+22.98*	*-1.10*	*+0.63*	*+1.14*	*+7.18*	*-3.51*
Bi-GRU	UTAD (MSE)	4.09	2.44	82.43	57.17	68.39	4.05
	STAD (BCE)	14.03	2.87	85.79	66.71	81.80	3.21
	Gain (Δ)	*+9.94*	*+0.43*	*+3.36*	*+9.54*	*+13.41*	*-0.84*

References

[1] J. Paparrizos, Y. Kang, P. Boniol, R. Tsay, T. Palpanas, M. J. Franklin, Tsb-uad: An end-to-end benchmark suite for univariate time-series anomaly detection, Proc. VLDB Endow. 15 (2022) 1697–1711.

[2] X. Wu, X. Qiu, Z. Li, Y. Wang, J. Hu, C. Guo, H. Xiong, B. Yang, CATCH: Channel-aware multivariate time series anomaly detection via frequency patching, in: International Conference on Learning Representations (ICLR), 2025, pp. 1–16.

[3] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, M. Kloft, Deep semi-supervised anomaly detection, in: International Conference on Learning Representations (ICLR), 2020, pp. 1–15.

[4] G. Pang, C. Shen, H. Jin, A. van den Hengel, Deep weakly-supervised anomaly detection, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2023, pp. 1795–1807. doi:10.1145/3580305.3599302.

[5] X. Zhou, Y. Wang, H. Xu, W. Liang, X. Li, G. Pang, IMPACT: Influence modeling for open-set time series anomaly detection, in: Proceedings of the 41st International Conference on Machine Learning (ICML), Vol. 235 of Proceedings of Machine Learning Research, 2024, pp. 62283–62300.

[6] M. C. du Plessis, G. Niu, M. Sugiyama, Analysis of learning from positive and unlabeled data, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 27, 2014, pp. 703–711.

[7] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6479–6488. doi:10.1109/CVPR.2018.00678.

[8] M. A. Siddiqui, A. Fern, T. G. Dietterich, R. Wright, A. Theriault, D. W. Archer, Feedback-guided anomaly discovery via online optimization, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2018, pp. 2200–2209. doi:10.1145/3219819.3220083.

[9] K. Ding, Q. Zhou, J. Li, K. Shu, H. Liu, Few-shot network anomaly detection via cross-network meta-learning, in: Proceedings of the Web Conference (WWW), 2021, pp. 2448–2456. doi:10.1145/3442381.3449922.

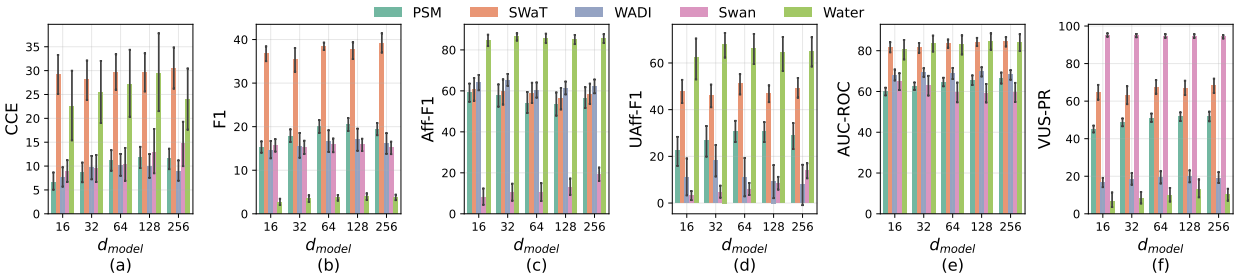


Figure S1: Sensitivity analysis of d_{model} across 5 datasets on six key performance metrics. (a) CCE, (b) F1 score, (c) Aff-F1 score, (d) UAff-F1 score, (e) AUC-ROC, (f) VUS-PR.

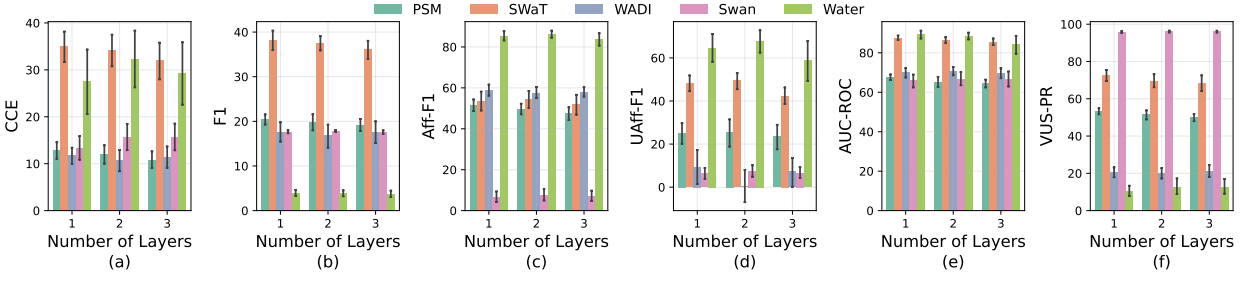


Figure S2: Sensitivity analysis of number of TEM layers across 5 datasets on six key performance metrics. (a) CCE, (b) F1 score, (c) Aff-F1 score, (d) UAff-F1 score, (e) AUC-ROC, (f) VUS-PR.

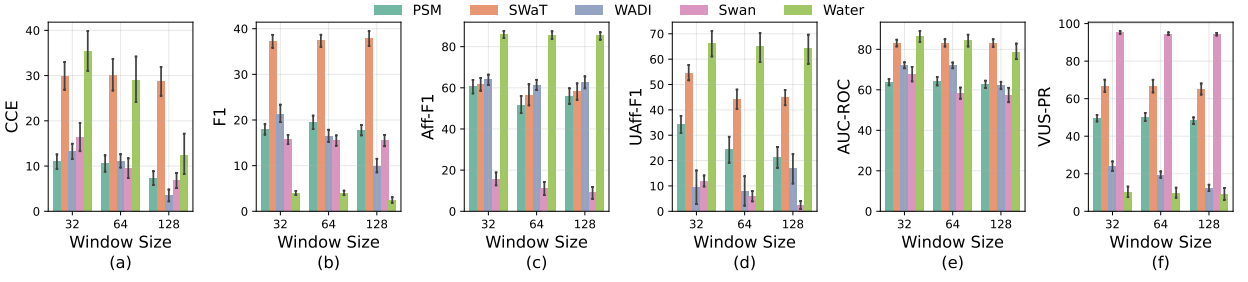


Figure S3: Sensitivity analysis of window size across 5 datasets on six key performance metrics. (a) CCE, (b) F1 score, (c) Aff-F1 score, (d) UAff-F1 score, (e) AUC-ROC, (f) VUS-PR.

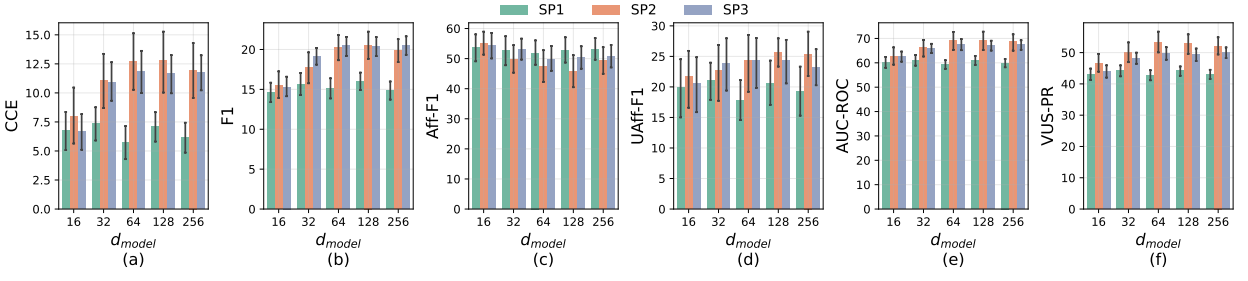


Figure S4: Sensitivity analysis of model dimension d_{model} across 3 subsets of the PSM dataset on six key performance metrics. (a) CCE, (b) F1 score, (c) Aff-F1 score, (d) UAff-F1 score, (e) AUC-ROC, (f) VUS-PR.

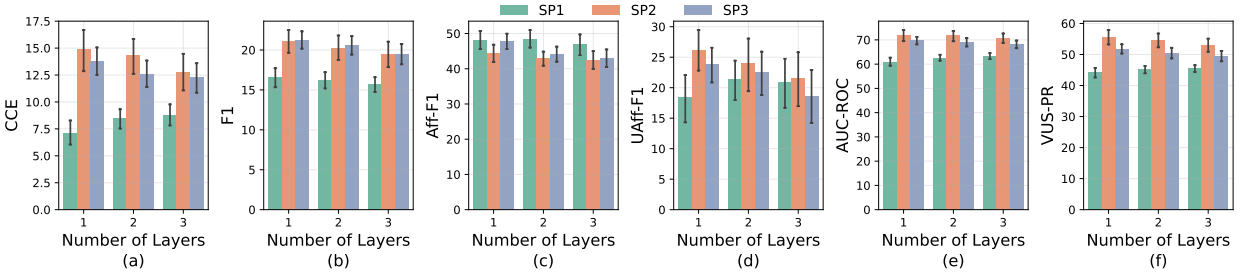


Figure S5: Sensitivity analysis of number of TEM layers across 3 subsets of the PSM dataset on six key performance metrics. (a) CCE, (b) F1 score, (c) Aff-F1 score, (d) UAff-F1 score, (e) AUC-ROC, (f) VUS-PR.

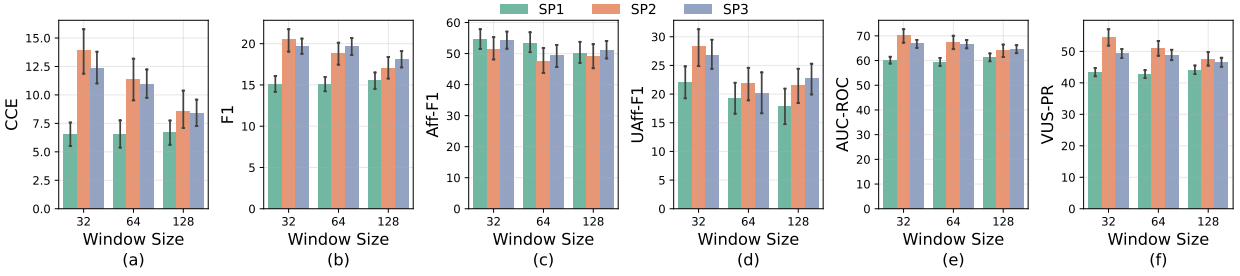


Figure S6: Sensitivity analysis of window size across 3 subsets of the PSM dataset on six key performance metrics. (a) CCE, (b) F1 score, (c) Aff-F1 score, (d) UAff-F1 score, (e) AUC-ROC, (f) VUS-PR.

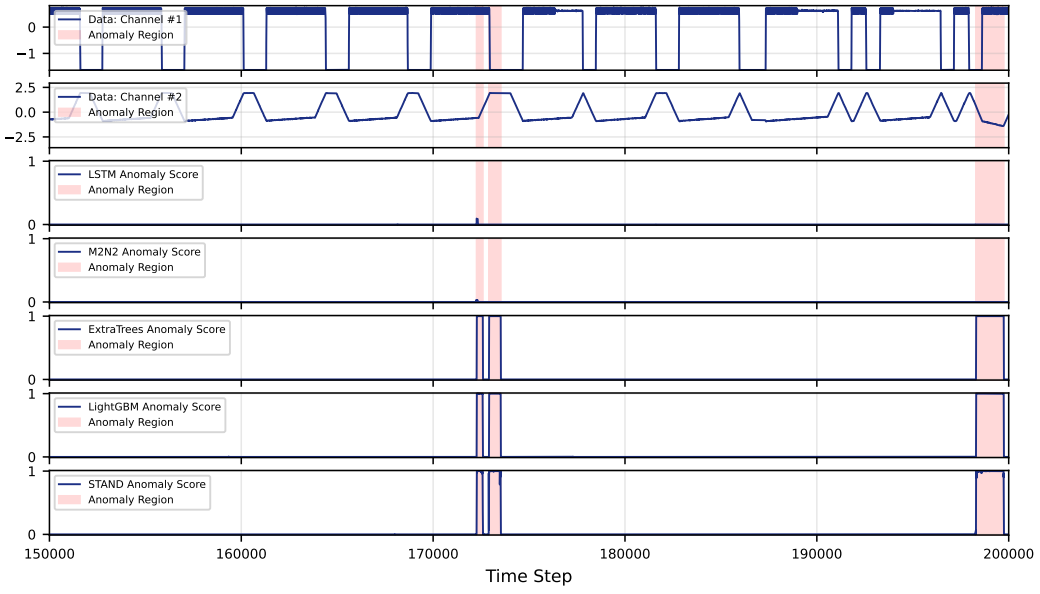


Figure S7: Anomaly score trajectories of representative models on a segment of the training prefix ($t \in [150\text{k}, 200\text{k}]$) on the SWaT benchmark. Supervised models fit the annotated anomalies cleanly, while unsupervised baselines fail to recognize these anomalous events.

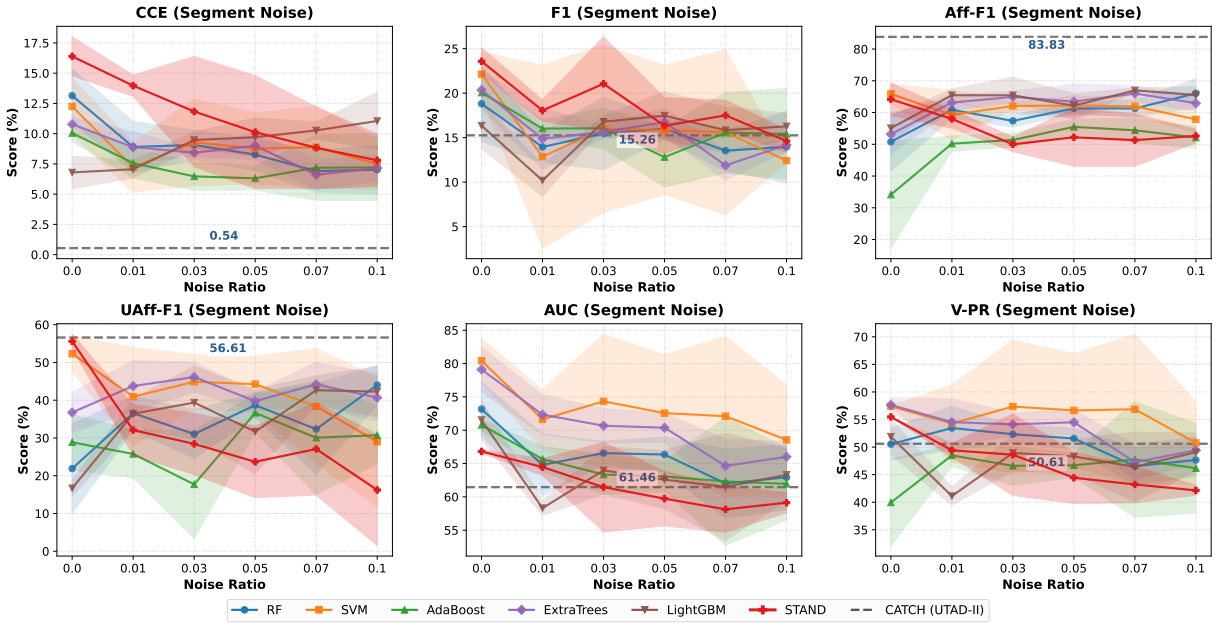


Figure S8: Robustness of supervised anomaly detectors under varying ratios of contiguous Segment Noise on the PSM benchmark. Solid curves depict the empirical mean across multiple repeated runs with distinct random corruption realizations, and shaded error bands denote the standard deviation ($\pm 1\sigma$). The horizontal dashed line indicates the performance of CATCH [2], a representative unsupervised UTAD-II baseline evaluated on the nominal sequence.