WHERE CAN NORMALITY GO: LEARNING ADMISSIBLE FUTURES FOR TIME SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series anomaly detection commonly uses prediction errors or likelihood-based scores to identify deviations from normal temporal evolution. However, even under the same observed history, normal evolution can follow multiple future trajectories, so deviation from a single expected trajectory need not indicate an anomaly. The set of normal continuations also varies with history: the same future trajectory may be admissible after one historical state but inconsistent with another. This motivates representing temporal normality as a history-conditioned region of admissible future trajectories and detecting anomalies when an observed future falls outside it. Learning such a region is challenging because each training history reveals only one realized normal continuation, providing a sample within the region without directly revealing its boundary. To address this, we propose **FutureBound**, a framework that learns this boundary implicitly through local future-space exploration and selective adversarial optimization. A context-conditioned generator produces multiple locally anchored yet diverse trajectories as probes around observed normal continuations. A discriminator then learns context-future compatibility by contrasting the observed continuation with nearby probes under the same history. To refine this distinction, the generator is adversarially optimized to make selected probes increasingly challenging to the discriminator, with proximity-based and direction-aware selection concentrating these updates around normal evolution. This interaction learns an implicit, potentially asymmetric boundary without prescribing its geometry in advance. At inference time, the learned compatibility directly measures whether an observed future agrees with its preceding history. Extensive experiments on four real-world time-series benchmarks demonstrate consistent improvements over existing methods, supporting history-conditioned future admissibility as a flexible reference for distinguishing anomalous deviations from normal temporal variation.

1 INTRODUCTION

Time-series anomaly detection (TSAD) (Blázquez-García et al., 2021; Schmidl et al., 2022) aims to identify observations that deviate from the normal evolution of a temporal system, with applications in industrial monitoring (Su et al., 2019), cyber-physical systems (Feng & Tian, 2021), and cloud services (Ren et al., 2019). Existing methods characterize normal temporal behavior through forecasting, reconstruction, or probabilistic modeling, and identify anomalies using the resulting errors or likelihood-based scores (Deng & Hooi, 2021; Li et al., 2019; Geiger et al., 2020; Audibert et al., 2020). The effectiveness of these scores depends on how well the learned reference captures the variations that can occur during normal evolution.

Normal future evolution, however, is not confined to a single expected trajectory. Even under the same observed history, temporal innovations can lead to different legitimate continuations (Wold, 1938). This variability persists even for an optimal deterministic predictor (Proposition 1), so prediction error reflects both normal variation and anomalous deviations. The range of normal continuations also changes with history: a rise in sensor readings may be expected during system startup but anomalous after a stable operating period. Building on boundary-based descriptions of normal-

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.

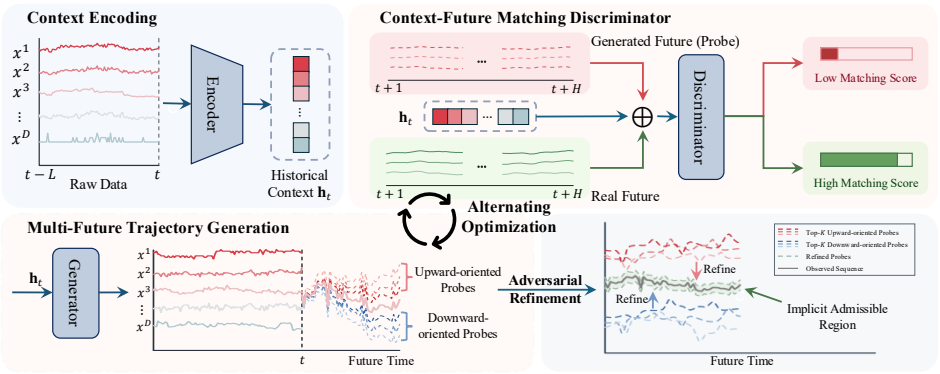


Figure 1: Overall architecture of FutureBound. The encoded historical context is shared by the generator and discriminator. Generated probes and the observed future are paired with the same context for matching evaluation, while selected probes from different residual orientations are progressively refined through alternating generator-discriminator optimization. The resulting trajectories delineate an implicit, potentially asymmetric admissible region around normal future evolution.

ity (Ruff et al., 2018; Goyal et al., 2020), *we represent temporal normality as a history-conditioned region of admissible future trajectories*.

Representing normality as a region also raises the question of its geometry. A likelihood threshold under a single Gaussian, for example, yields an ellipsoid symmetric about the conditional mean; learning its parameters does not remove this geometric restriction. When normal futures occupy separate modes, such a region can include unsupported trajectories between them and miss corresponding anomalies (Proposition 2). We therefore seek a region whose shape is learned from normal temporal behavior, adapts to history, and allows asymmetric variations.

Learning this region from normal-only data presents a further challenge. Each historical window is paired with only one realized normal continuation, which provides a sample within the admissible set but does not reveal where its boundary lies. Our key insight is that *an observed normal continuation can anchor local exploration of the future space, providing comparisons for learning the surrounding boundary*.

Building on this insight, we propose **FutureBound**, which learns an implicit history-conditioned boundary through local future-space exploration and selective adversarial optimization (Fig. 1). Given the historical context, a generator produces multiple candidate trajectories as probes around the observed normal continuation. Fitting objectives anchor these probes to normal evolution in both values and temporal changes, while a diversity regularizer encourages them to explore different nearby alternatives. A discriminator then learns context-future compatibility by comparing the observed continuation with selected nearby probes under the same history, assigning higher scores to the observed normal future. To refine this distinction, we alternate discriminator training with adversarial generator updates that make selected probes increasingly difficult to distinguish from observed normal continuations. Proximity-based selection focuses these updates on nearby trajectories, while direction-aware selection retains different residual orientations to account for potentially asymmetric local boundaries. This interaction progressively refines the learned boundary without prescribing its geometry. At inference time, the discriminator directly evaluates whether an observed future agrees with its preceding history, with low compatibility indicating an anomaly. Extensive experiments on four real-world multivariate benchmarks demonstrate that FutureBound consistently achieves strong detection performance, while qualitative and ablation analyses further verify the effectiveness of its learned future boundaries.

Our main contributions are as follows:

- We formulate temporal normality as a history-conditioned region of admissible future trajectories, making the relationship between history and future the basis for anomaly detection.
- We introduce local future-space exploration with a context-conditioned generator, producing locally anchored yet diverse probes for comparisons beyond a single observed continuation.

- We develop selective adversarial boundary learning guided by proximity and residual orientation, enabling history-dependent decision regions with flexible, potentially asymmetric geometry.

2 METHODOLOGY

2.1 PROBLEM FORMULATION AND MOTIVATION

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ denote a multivariate time series, where $\mathbf{x}_t \in \mathbb{R}^D$ contains the observations of D variables at timestamp t. Under the unsupervised setting, training data are assumed to contain normal behavior, while anomalies are present only during testing phase. For each timestamp t, we define the preceding L observations as the historical context $\mathbf{H}_t = [\mathbf{x}_{t-L}, \ldots, \mathbf{x}_{t-1}] \in \mathbb{R}^{L\times D}$, and the subsequent H observations as the future trajectory $\mathbf{F}_t = [\mathbf{x}_t, \cdots, \mathbf{x}_{t+H-1}] \in \mathbb{R}^{H\times D}$.

However, we argue that normal future evolution is not necessarily confined to a single trajectory. Even under the same historical context, temporal innovations may lead to different legitimate continuations. From a prediction perspective, the conditional expectation provides the optimal deterministic prediction under the mean-squared-error criterion, while temporal innovations characterize the variation that remains unpredictable from the observed history (Wold, 1938). Proposition 1 (with proof in Appendix A) formalizes the resulting irreducible variability of normal future trajectories.

Proposition 1. *Given the historical context $\mathbf{H}_t$ and future trajectory $\mathbf{F}_t$ defined above, suppose the first future observation follows*

$$\mathbf{x}_t = g(\mathbf{H}_t) + \boldsymbol{\varepsilon}_t,$$

where $g(\cdot)$ describes the predictable component of the temporal evolution and $\boldsymbol{\varepsilon}_t \in \mathbb{R}^D$ denotes the innovation satisfying $\mathbb{E}_{\mathrm{N}}[\boldsymbol{\varepsilon}_t|\mathbf{H}_t] = \mathbf{0}$ and $\mathbb{E}_{\mathrm{N}}\left[\|\boldsymbol{\varepsilon}_t\|_2^2\right] > 0$. Let $\boldsymbol{\mu}_t = \mathbb{E}_{\mathrm{N}}[\mathbf{F}_t|\mathbf{H}_t]$ represent the conditional mean future trajectory, and define the conditional future variability as $V_t = \mathbb{E}_{\mathrm{N}}\left[\|\mathbf{F}_t - \boldsymbol{\mu}_t\|_F^2|\mathbf{H}_t\right]$. Then

$$\mathbb{E}_{\mathrm{N}}[V_t] \geq \mathbb{E}_{\mathrm{N}}\left[\|\boldsymbol{\varepsilon}_t\|_2^2\right] > 0.$$

Consequently, for any deterministic future predictor $f(\cdot)$,

$$\inf_f \mathbb{E}_{\mathrm{N}}\left[\|\mathbf{F}_t - f(\mathbf{H}_t)\|_F^2\right] = \mathbb{E}_{\mathrm{N}}[V_t] > 0.$$

This proposition shows that even the optimal deterministic predictor retains a nonzero error on normal future trajectories, since part of the discrepancy arises from intrinsic temporal variation. A single predicted trajectory therefore provides an incomplete description of normal future evolution. Instead, normality should account for a set of legitimate continuations under the same historical context, motivating an admissible future region rather than a single expected trajectory.

Yet a region with an inappropriate geometry may overextend to include trajectories that are unsupported by the underlying temporal evolution. A Gaussian assumption provides a simple example in which the imposed region may regard trajectories that are not actually admissible as highly normal.

Example 1. *Consider a fixed historical context $\mathbf{H}$ under which two separated future trajectories $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ are both normal continuations. Assume that they occur symmetrically around their conditional mean $\bar{\mathbf{F}} = (\mathbf{F}^{(1)} + \mathbf{F}^{(2)})/2$, while the intermediate trajectory $\bar{\mathbf{F}}$ itself is not supported by the normal evolution. Under the squared prediction loss, $\bar{\mathbf{F}}$ is still the optimal deterministic prediction. Moreover, if these normal futures are approximated by a single Gaussian distribution, $\bar{\mathbf{F}}$ becomes the center of the resulting Gaussian region and receives the highest Gaussian density. Consequently, a future trajectory that is not admissible under the actual temporal evolution can nevertheless be regarded as highly normal under the imposed Gaussian assumption.*

The example above points to a more general problem: A predefined distribution may impose a geometry that does not match the actual structure of admissible future evolution. Under the Gaussian assumption, admissible futures are represented by a convex ellipsoidal region centered at the conditional mean. When the true admissible future space is non-convex, such a region may extend beyond the supported normal evolution and consequently include inadmissible trajectories.

We also investigate this issue through an empirical analysis of prediction residuals on four real-world benchmarks in Fig. 2. After sensor-wise standardization, a Gaussian distribution is fitted to the normal calibration residuals for each variable and future step. The fitted distribution is then evaluated against both normal and anomalous test residuals using the Kolmogorov-Smirnov (KS) statistic (Smirnov, 1948). It reveals substantial Gaussian mismatch for normal residuals, while anomalous residuals exhibit considerable overlap with the normal KS distributions. These results suggest that a fixed Gaussian form may neither reliably characterize normal future variation nor clearly distinguish it from anomalous behavior. More generally, this limitation extends to convex decision regions, as formalized below.

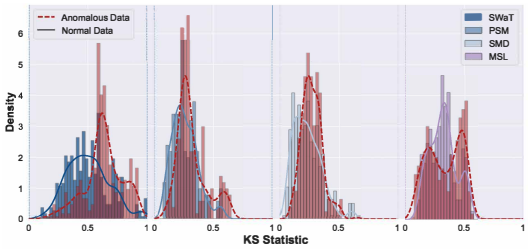


Figure 2: Gaussianity diagnostics of prediction residuals across four real-world datasets.

Proposition 2. *For a fixed historical context* $\mathbf{H}$*, let* $\mathcal{F}^*(\mathbf{H})$ *denote the admissible normal future region. Suppose* $\mathcal{F}^*(\mathbf{H})$ *is non-convex, such that there exist two admissible future trajectories* $\mathbf{F}^{(1)}, \mathbf{F}^{(2)} \in \mathcal{F}^*(\mathbf{H})$ *and some* $\lambda \in (0,1)$ *satisfying*

$$\lambda \mathbf{F}^{(1)} + (1-\lambda)\mathbf{F}^{(2)} \notin \mathcal{F}^*(\mathbf{H}).$$

Then any convex region $\mathcal{C}(\mathbf{H})$ *containing both* $\mathbf{F}^{(1)}$ *and* $\mathbf{F}^{(2)}$ *must also contain* $\lambda\mathbf{F}^{(1)} + (1-\lambda)\mathbf{F}^{(2)}$*. Consequently,*

$$\mathcal{C}(\mathbf{H}) \setminus \mathcal{F}^*(\mathbf{H}) \neq \varnothing.$$

Therefore, a single convex decision region cannot exactly represent a non-convex admissible future space without including unsupported trajectories.

Taken together, these observations motivate treating admissible future evolution as a context-specific region whose form must be inferred from normal temporal behavior. Accordingly, let $\mathcal{F}^*(\mathbf{H}_t)$ denote the unknown region of admissible normal future trajectories associated with the historical context $\mathbf{H}_t$. A meaningful region should cover most normal continuations while remaining sufficiently tight. Otherwise, high coverage could be achieved simply by expanding the region, reducing its ability to distinguish normal from abnormal evolution. This leads to the following objective:

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{H}\sim P_{\mathrm{N}}(\mathbf{H})}\left[\Omega(\mathcal{F}(\mathbf{H}))\right] \quad \text{s.t.} \quad \mathbb{P}_{(\mathbf{H},\mathbf{F})\sim P_{\mathrm{N}}}\left[\mathbf{F} \in \mathcal{F}(\mathbf{H})\right] \geq 1-\delta, \tag{1}$$

where P_{N} denotes the joint distribution of historical contexts and their subsequent normal trajectories, and $P_{\mathrm{N}}(\mathbf{H})$ is its marginal distribution over historical contexts. $\Omega(\cdot)$ measures the extent of a future region, while $\delta \in (0,1)$ specifies the tolerated normal miscoverage rate. In practice, however, both P_{N} and the geometry of $\mathcal{F}^*(\mathbf{H})$ are unknown, making the admissible region difficult to recover directly in the high-dimensional trajectory space. To circumvent explicit recovery, the proposed **FutureBound** explores the future space around observed normal continuations and learns the surrogate admissible boundary implicitly. The following subsections detail the formulation and optimization of its individual components.

2.2 Contextual Future Space Exploration

Given the historical context $\mathbf{H}_t$, we first encode its preceding temporal evolution into a compact representation using a temporal convolutional network (TCN) (Bai et al., 2018). Specifically, the historical window is arranged as $\mathbf{Z}_t^{(0)} = \mathbf{H}_t^\top \in \mathbb{R}^{D\times L}$, with individual variables treated as input channels along the temporal dimension. The context representation is obtained as

$$\mathbf{h}_t = \mathrm{TCN}_\phi\left(\mathbf{H}_t^\top\right)_{:,L}, \tag{2}$$

where ϕ denotes the encoder parameters. The resulting representation $\mathbf{h}_t$ summarizes the preceding temporal context and conditions the subsequent future-space exploration and boundary learning.

Multi-Future Trajectory Generation. For each historical context $\mathbf{H}_t$, the training data reveal only one realized future $\mathbf{F}_t$, although multiple continuations may be admissible under the same

context. A single realization therefore provides only a sparse observation of the surrounding future space. To probe this unobserved space, we use the context representation $\mathbf{h}_t$ to jointly generate M context-conditioned candidate trajectories:

$$\left[\hat{\mathbf{F}}_t^{(1)}, \hat{\mathbf{F}}_t^{(2)}, \cdots, \hat{\mathbf{F}}_t^{(M)}\right] = \mathcal{G}_\theta(\mathbf{h}_t), \tag{3}$$

where θ denotes the generator parameters and M is the number of generated candidates. These candidates are not assumed to be normal futures, but instead serve as local probes around the observed continuation. To keep the probes close to the observed normal evolution, we measure the discrepancy of each candidate at both the state and transition levels:

$$d_t^{(m)} = \frac{1}{HD}\left\|\mathbf{F}_t - \hat{\mathbf{F}}_t^{(m)}\right\|_F^2 + \frac{\lambda_\Delta}{(H-1)D}\left\|\Delta\mathbf{F}_t - \Delta\hat{\mathbf{F}}_t^{(m)}\right\|_F^2, \tag{4}$$

where $\Delta\mathbf{F}_t \in \mathbb{R}^{(H-1)\times D}$ contains the differences between consecutive states and λ_Δ controls the contribution of the transition discrepancy. The state discrepancy captures differences in future values, while the transition discrepancy reflects deviations in temporal evolution patterns.

Let $\mathcal{S}_{\text{train}} = \{(\mathbf{H}_t, \mathbf{F}_t)\}_{t\in\mathcal{T}_{\text{train}}}$ denote the training set, where $\mathcal{T}_{\text{train}}$ contains all training anchor timestamps. The fitting objective over all generated candidates is defined as

$$\mathcal{L}_{\text{fit}} = \frac{1}{|\mathcal{S}_{\text{train}}|M}\sum_{t\in\mathcal{T}_{\text{train}}}\sum_{m=1}^{M} d_t^{(m)}. \tag{5}$$

Applying this objective to every candidate keeps the generated probes locally anchored around the observed normal continuation and prevents them from drifting arbitrarily in the trajectory space.

Diversity Regularizer. Anchoring all candidates to the same observed future may pull them toward similar trajectories, weakening their ability to probe the surrounding space. To discourage such concentration, we introduce a diversity regularizer based on the pairwise trajectory discrepancy $\delta_t^{(m,n)} = \frac{1}{HD}\|\hat{\mathbf{F}}_t^{(m)} - \hat{\mathbf{F}}_t^{(n)}\|_F^2$:

$$\mathcal{L}_{\text{div}} = \frac{1}{|\mathcal{S}_{\text{train}}|M(M-1)}\sum_{t\in\mathcal{T}_{\text{train}}}\sum_{m\neq n}\exp\left(-\frac{\delta_t^{(m,n)}}{\sigma_{\text{div}}}\right), \tag{6}$$

where σ_{div} controls the sensitivity of the penalty to pairwise discrepancy. Minimizing $\mathcal{L}_{\text{div}}$ discourages candidate trajectories from collapsing to similar solutions, while $\mathcal{L}_{\text{fit}}$ prevents excessive dispersion away from the observed normal continuation. Together, the two objectives produce locally anchored yet diverse probes for subsequent boundary learning.

2.3 Boundary-Oriented Adversarial Learning

Context-Conditioned Discriminator. By treating the generated trajectories as local probes around the observed normal future, we introduce a context-conditioned discriminator $\mathcal{D}_\psi(\mathbf{h}_t, \mathbf{F})$ to learn the admissible boundary, where a larger value indicates stronger compatibility between the historical context and the future trajectory. To concentrate boundary learning around normal evolution, we select the $K_{\mathcal{D}}$ candidates with the smallest $d_t^{(m)}$ and denote their index set by $\mathcal{I}_t^{\mathcal{D}}$. The discriminator is optimized by

$$\mathcal{L}_{\text{dis}} = \frac{1}{|\mathcal{S}_{\text{train}}|K_{\mathcal{D}}}\sum_{t\in\mathcal{T}_{\text{train}}}\sum_{m\in\mathcal{I}_t^{\mathcal{D}}}\mathcal{D}_\psi\left(\mathbf{h}_t, \hat{\mathbf{F}}_t^{(m)}\right) - \frac{1}{|\mathcal{S}_{\text{train}}|}\sum_{t\in\mathcal{T}_{\text{train}}}\mathcal{D}_\psi(\mathbf{h}_t, \mathbf{F}_t). \tag{7}$$

Minimizing $\mathcal{L}_{\text{dis}}$ encourages the observed normal future to receive a higher compatibility score than its nearby probes. Focusing on the nearest probes avoids trivial discrimination against distant trajectories and directs learning toward the local admissible boundary.

Selective Adversarial Optimization. The fitting and diversity objectives provide locally distributed probes, while boundary refinement relies on making these probes increasingly challenging to the discriminator. With the discriminator fixed, the generator is adversarially optimized so that

selected candidates receive higher compatibility scores, forcing the discriminator to resolve progressively finer distinctions around normal evolution. This adversarial update is restricted to a small subset of nearby probes, since distant candidates are less relevant to local boundary refinement while increasing the optimization cost. The selection criterion derives from the trajectory discrepancy $d_t^{(m)}$ in Eq. (4), keeping probe selection anchored to the observed normal continuation and independent of the discriminator's current responses. However, $d_t^{(m)}$ captures only the magnitude of trajectory discrepancy and discards its **orientation**. This distinction matters because the admissible boundary need not be symmetric around the observed future; probes with comparable discrepancies but opposite deviations may therefore reveal different local boundary characteristics. Here we define a residual orientation score to supplement the proximity criterion with this directional information:

$$c_t^{(m)} = \frac{1}{HD} \sum_{i=1}^{H} \sum_{j=1}^{D} \left(\hat{\mathbf{F}}_{t,i,j}^{(m)} - \mathbf{F}_{t,i,j} \right). \tag{8}$$

Nonnegative and negative values of $c_t^{(m)}$ partition the candidates into two orientation groups, $\mathcal{U}_t = \{m : c_t^{(m)} \geq 0\}$ and $\mathcal{L}_t = \{m : c_t^{(m)} < 0\}$. Within each group, the $K_{\mathcal{G}}$ candidates with the smallest trajectory discrepancy $d_t^{(m)}$ are retained, and their union forms the adversarial probe set $\mathcal{I}_t^{\mathcal{G}}$. If either group contains fewer than $K_{\mathcal{G}}$ candidates, the remaining positions are filled by the nearest unused candidates from the complete candidate set. This selection preserves local proximity while accounting for the potentially asymmetric boundary structure around the observed future.

During generator optimization, the discriminator parameters remain fixed, and the selected probes are encouraged to receive higher compatibility scores through:

$$\mathcal{L}_{\text{adv}} = -\frac{1}{|\mathcal{S}_{\text{train}}|} \sum_{t \in \mathcal{T}_{\text{train}}} \frac{1}{|\mathcal{I}_t^{\mathcal{G}}|} \sum_{m \in \mathcal{I}_t^{\mathcal{G}}} \mathcal{D}_\psi \left(\mathbf{h}_t, \hat{\mathbf{F}}_t^{(m)} \right). \tag{9}$$

By raising the compatibility scores of selected probes, the adversarial objective makes them more challenging for the discriminator and promotes finer boundary refinement in subsequent updates.

Then the overall generator objective can be formulated as

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\text{adv}} + \alpha \mathcal{L}_{\text{fit}} + \beta \mathcal{L}_{\text{div}}, \tag{10}$$

where α and β balance local anchoring and trajectory diversity, respectively. The three terms jointly maintain proximity to observed normal evolution, preserve exploration across the local future space, and expose the discriminator to increasingly challenging probes. Alternating optimization of the generator and discriminator progressively refines the context-dependent scoring function around normal future evolution, yielding an implicit admissible boundary.

Anomaly Score. The learned discriminator should provide a context-dependent compatibility score for any observed future trajectory. Since a larger output indicates stronger agreement between the future trajectory and its preceding history, we first define the trajectory-level anomaly score as

$$a_t = -\mathcal{D}_\psi(\mathbf{h}_t, \mathbf{F}_t). \tag{11}$$

Then let $\mathcal{W}(\tau) = \{t : t \leq \tau \leq t + H - 1\}$ denote the set of future trajectories covering timestamp τ, and $k_\tau = \min(k, |\mathcal{W}(\tau)|)$, the timestamp-level anomaly score is obtained by averaging the largest k_τ trajectory-level scores:

$$s_\tau = \frac{1}{k_\tau} \sum_{a_t \in \text{Top}_{k_\tau}\{a_t : t \in \mathcal{W}(\tau)\}} a_t. \tag{12}$$

A larger s_τ indicates stronger inconsistency between the observed evolution around timestamp τ and the learned context-dependent normal behavior.

3 Connection with Previous Work

FutureBound builds upon several fundamental ideas in time series anomaly detection, including history-conditioned prediction, adversarial generation, and boundary learning, while differing in

how these ideas are integrated to define normality. **1)** Similar to prediction-based methods (Zhao et al., 2020; Deng & Hooi, 2021), FutureBound leverages historical contexts to characterize future evolution. Here we would like to emphasize that probabilistic forecasters such as TimeGrad (Rasul et al., 2021) already account for multiple possible futures; however, their objective is conditional predictive distribution estimation, and applying them to anomaly detection still requires an additional scoring criterion, such as likelihood, sample distance, or denoising error. FutureBound instead directly learns history-future compatibility for anomaly detection, with generated trajectories serving only as local probes for refining the admissible boundary. **2)** FutureBound also adopts an adversarial generation framework related to GAN-based anomaly detectors such as MAD-GAN (Li et al., 2019) and TadGAN (Geiger et al., 2020); however, generated trajectories are not optimized to reproduce realistic samples from a global normal distribution, but serve as local probes for future-space exploration and boundary refinement. **3)** Unlike boundary-based anomaly detection methods such as Deep SVDD (Ruff et al., 2018) and DROCC (Goyal et al., 2020), which define normal regions in fixed latent representation spaces, FutureBound learns a history-dependent future boundary, where each historical context induces its own admissible future region. Taken together, FutureBound does not merely combine prediction, generation, and boundary learning, but shifts the object of time series anomaly detection from global sequence normality to context-dependent future admissibility. Detailed related work can be found in Appendix I.

4 Experiment

4.1 Experimental Settings

Datasets. Evaluation is performed on four multivariate TSAD datasets from three application scenarios: industrial control (SWaT), cloud and server monitoring (PSM and SMD), and spacecraft telemetry (MSL). Their main statistics are listed in Appendix D.

Baselines. FutureBound is compared with representative methods from five major multivariate TSAD paradigms: **(1) Classical Detectors:** k-NN (Ramaswamy et al., 2000), OCSVM (Schölkopf et al., 1999), LOF (Breunig et al., 2000), IForest (Liu et al., 2008), and COPOD (Li et al., 2020); **(2) Geometric Boundary-based Methods:** Deep SVDD (Ruff et al., 2018), COUTA (Xu et al., 2024), and GBOC (Shen et al., 2026); **(3) Reconstruction- and Prediction-based Methods:** USAD (Audibert et al., 2020), TcnED (Garg et al., 2021), TranAD (Tuli et al., 2022), TimesNet (Wu et al., 2023), SDA2D (Sun et al., 2026), and GCAD (Liu et al., 2025); **(4) Representation- and Discrepancy-based Methods:** AnomalyTrans (Xu et al., 2022), DCdetector (Yang et al., 2023), and PaAno (Park & Kang, 2026); and **(5) Synthetic-anomaly-based Methods:** MAD-GAN (Li et al., 2019), NCAD (Carmona et al., 2022), and SPAGD (Cai et al., 2025).

Implementation Details. Detailed implementation and hyper-parameter settings are provided in Appendix E. Performance is evaluated using AUC-ROC, AUPRC, and F1-Score. Point adjustment is excluded from all evaluations.

4.2 Numerical Results

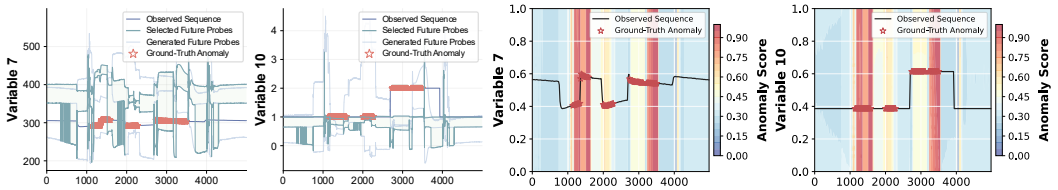
FutureBound achieves the best overall performance on SWaT, PSM, MSL, and SMD across AUC, AUPRC, and F1-Score, as shown in Table 1. While competing methods characterize normality through different references, their performance varies noticeably across datasets with different temporal dynamics. This suggests that a single global reference or fixed normality description may not transfer equally well across operating regimes. FutureBound instead evaluates each observed future relative to its own preceding history and explores the corresponding local future space with multiple generated probes. The discriminator then learns whether a candidate continuation remains compatible with that specific context, allowing the effective decision region to adapt across histories rather than being shared globally.

4.3 Qualitative Analysis

Decision Boundary Visualizations. Fig. 3 visualizes the future probes generated from randomly sampled 5,000 timestamps and the discriminator responses on 2 representative SWaT variables, with

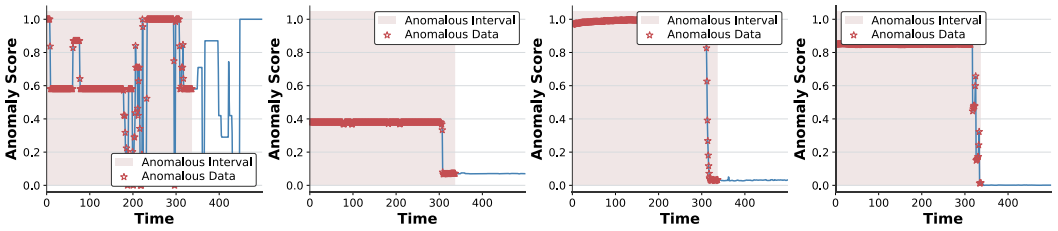
Table 1: Average AUCs, AUPRCs and F1-Scores with standard deviation (10 trials) of different multivariate time series anomaly detection algorithms. The best results are **bolded**.

Method	SWaT			PSM			MSL			SMD		
	AUC	AUPRC	F1-Score	AUC	AUPRC	F1-Score	AUC	AUPRC	F1-Score	AUC	AUPRC	F1-Score
k-NN	82.33±0.00	72.58±0.00	76.80±0.00	63.64±0.00	46.89±0.00	46.44±0.00	49.38±0.00	12.70±0.00	19.17±0.00	64.07±0.00	12.70±0.00	21.94±0.00
OCSVM	80.59±0.00	70.41±0.00	73.53±0.00	47.48±0.00	25.56±0.00	44.06±0.00	59.61±0.00	20.29±0.00	25.58±0.00	52.48±0.00	5.12±0.00	9.27±0.00
LOF	80.73±0.00	70.93±0.00	75.65±0.00	61.66±0.00	45.30±0.00	43.76±0.00	40.43±0.00	8.50±0.00	19.06±0.00	49.07±0.00	4.88±0.00	8.44±0.00
IForest	84.18±0.39	74.62±0.50	76.22±0.92	66.88±0.71	39.16±1.04	50.48±0.75	57.95±0.79	13.92±0.78	22.27±0.36	66.33±0.76	12.42±0.84	15.56±0.47
Deep SVDD	68.64±14.37	34.35±23.11	50.06±18.67	52.32±8.42	35.11±9.72	44.41±1.29	57.70±5.25	8.05±1.77	13.31±2.67	54.51±1.26	14.51±0.04	20.95±0.66
MAD-GAN	83.78±0.53	74.21±0.45	75.89±0.10	61.99±1.84	35.42±1.18	48.08±2.48	54.06±1.77	13.79±0.54	20.00±0.77	64.90±1.13	11.14±1.79	17.56±3.68
COPOD	81.66±0.00	71.57±0.00	70.85±0.00	65.12±0.00	39.93±0.00	48.80±0.00	57.52±0.00	14.38±0.00	23.96±0.00	65.44±0.00	12.40±0.00	16.69±0.00
USAD	82.18±0.15	72.76±0.08	76.57±0.00	63.91±0.08	43.18±0.10	47.41±0.11	55.31±0.70	14.75±0.09	21.23±0.19	66.40±0.56	12.80±0.26	22.06±0.16
TcnED	83.00±0.13	74.13±0.13	76.32±0.21	50.54±0.69	29.50±0.43	43.48±0.00	47.31±2.90	11.92±1.89	21.03±2.56	42.34±1.08	4.43±0.19	7.98±0.00
TranAD	81.72±0.20	72.44±0.21	76.66±0.05	66.58±0.70	46.62±0.67	48.90±0.70	51.78±1.47	13.58±0.36	22.52±1.42	71.16±0.52	12.71±0.16	20.59±0.28
AnomalyTrans	81.83±0.11	71.81±0.24	76.35±0.17	69.91±0.31	47.39±0.76	50.57±0.34	57.62±0.90	13.47±0.33	21.49±0.18	75.21±0.29	12.22±0.65	18.92±0.84
NCAD	81.16±2.29	53.54±4.19	70.56±1.16	60.94±5.39	37.88±2.49	46.18±3.63	60.67±1.53	15.06±0.35	22.77±0.77	60.71±0.44	7.81±0.28	13.80±0.44
COUTA	78.47±5.29	55.43±27.29	65.14±18.36	73.39±2.22	50.66±2.91	54.13±4.20	54.76±2.33	14.20±0.19	21.47±1.75	70.73±2.08	16.98±0.56	21.14±1.14
TimesNet	24.33±0.27	9.25±0.09	21.66±0.00	59.30±0.21	39.12±0.21	43.49±0.00	62.28±0.21	14.94±0.05	24.23±0.15	75.18±0.14	14.98±0.24	21.10±0.33
DCdetector	49.90±0.12	12.12±0.03	21.67±0.01	50.06±0.17	27.78±0.07	43.51±0.00	50.21±0.16	10.58±0.02	19.09±0.00	50.18±0.14	4.17±0.01	7.98±0.00
SPAGD	73.76±4.24	29.55±6.51	40.35±9.06	49.64±6.94	30.14±3.97	48.77±1.20	54.93±2.87	16.73±3.41	26.33±1.70	51.17±0.60	9.77±0.84	16.63±0.19
GCAD	45.39±1.73	11.03±0.19	23.15±0.69	68.34±0.08	46.29±0.03	48.51±0.08	57.93±2.34	15.38±0.59	23.77±1.25	70.08±0.61	15.13±0.50	19.47±1.48
SDA2D	82.03±0.04	72.22±0.23	76.07±0.06	68.56±0.49	50.16±0.43	47.97±0.31	54.61±0.22	13.62±0.48	19.73±0.12	70.86±0.21	14.05±0.21	21.84±0.28
GBOC	36.83±1.88	14.10±0.49	12.26±3.28	54.09±0.00	33.91±1.23	5.17±1.35	59.16±0.02	13.01±0.04	8.12±1.18	76.59±0.42	20.10±3.77	25.93±4.19
PaAno	27.29±2.84	8.20±0.09	23.36±1.10	67.27±3.20	41.42±2.84	50.67±2.93	70.98±1.91	20.97±0.98	28.90±1.97	77.02±1.50	10.73±0.91	20.00±0.76
Ours	**85.45±1.16**	**75.12±0.74**	**76.84±0.02**	**76.19±3.27**	**54.39±4.39**	**56.77±2.37**	**72.34±0.92**	**24.19±1.42**	**31.67±1.34**	**77.20±0.54**	**21.16±3.35**	**28.05±2.93**



(a) Generated Future Probe (b) Learned Boundary Response

Figure 3: Decision boundary visualizations of FutureBound on representative SWaT variables. (a) Generated and selected future probes on testing data; (b) Learned boundary responses corresponding to the observed testing sequences.



(a) LOF (b) SPAGD (c) GBOC (d) FutureBound

Figure 4: Anomaly score visualization comparison on a 500-point SWaT testing segment.

additional results of other variables on SWaT and PSM provided in Appendix F. In Subfigure (a), normal observations are largely covered by the generated trajectories, while anomalous segments deviate from the probe-supported region, showing that the generator provides a coarse local future space. The selected probes further concentrate around the observed evolution and provide local comparisons. In Subfigure (b), the discriminator refines this coarse separation: anomalous timestamps receive higher responses, while normal regions remain low. Compared with the probes, discriminator responses are more localized around anomalous intervals, indicating that probes explore the future space, whereas the discriminator forms the final context-dependent decision boundary.

Detection Behavior Comparison. We also visualize the timestamp-level anomaly scores of representative detectors on the same 500-point SWaT segment in Fig. 4. LOF produces fragmented activations, whereas SPAGD and GBOC show relatively cleaner separation between anomalous and normal regions. These two methods provide particularly relevant comparisons to FutureBound: SPAGD represents generative and synthetic-anomaly-based detection, while GBOC models normality using a predefined granular-ball geometric form. FutureBound maintains a stable high response throughout the anomalous interval and drops sharply once normal evolution resumes, yielding a more compact transition with little residual activation. This comparison highlights the benefit of learning a history-conditioned future boundary. The complete results are provided in Appendix F.

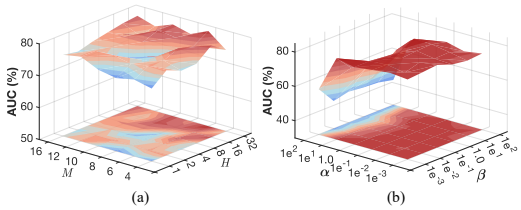


Figure 5: Parameter sensitivity on SMD under AUC. Sensitivity to (a) generated trajectories number M and future horizon H; (b) coefficients α and β of the fitting and diversity losses.

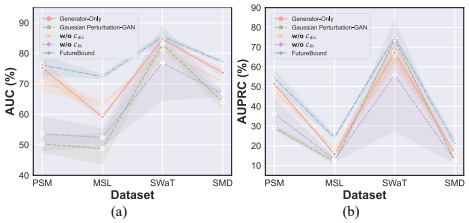


Figure 6: Ablation analyses on four datasets in terms of AUC and AUPRC, evaluating the effects of adversarial boundary learning, probe diversity, local fitting, and learned trajectory generation.

4.4 Parameter Sensitivity

We provide a sensitivity analysis of main hyper-parameters on SMD in Fig. 5. As shown in (a), increasing M or H does not monotonically improve detection: insufficient probes or a short horizon limits the exploration of alternative futures, whereas overly large settings provide little additional benefit and may introduce redundant or harder-to-model trajectories. Subfigure (b) further reveals the importance of balancing α and β. Extreme weighting of either local fitting or trajectory diversity degrades performance, while moderate combinations better preserve both probe locality and exploration. These results are consistent with the design of FutureBound, where effective boundary learning requires sufficient future-space exploration without losing the connection to normal evolution. Due to space limitations, we defer the complete parameter sensitivity results, together with additional analyses on PSM and SMD, to Appendix G.

4.5 Ablation Study

To assess the contribution of each component, we consider four variants: **1) Generator-Only**, which evaluates anomalies using a distance-based tube over generated probes; **2) w/o** $\mathcal{L}_{\text{div}}$, which removes diversity regularization and limits local future-space exploration; **3) w/o** $\mathcal{L}_{\text{fit}}$, which removes the fitting constraint anchoring probes to normal evolution; and **4) Gaussian Perturbation-GAN**, which replaces learned probes with Gaussian perturbations while retaining adversarial boundary learning.

The ablation results of Fig. 6 demonstrate the complementary roles of key components, with complete results reported in Appendix H. Generator-Only remains competitive on some datasets, but its weaker performance on SMD and MSL, together with the clear gains from adding the discriminator on MSL, indicates that generated trajectories alone provide only coarse future-space coverage and require adversarial learning for effective boundary refinement. Removing $\mathcal{L}_{\text{div}}$ consistently reduces performance, showing that insufficient probe diversity limits local future exploration. Removing $\mathcal{L}_{\text{fit}}$ causes larger degradation and higher variance, highlighting the importance of anchoring probes to normal evolution. Besides, Gaussian Perturbation-GAN performs inconsistently across datasets, suggesting that generic perturbations cannot reliably capture diverse temporal dynamics.

5 Conclusion

In this work, we revisit a fundamental limitation of time-series anomaly detection: normal future evolution cannot be fully represented by a single expected trajectory or a region with predefined geometry, since admissible futures vary with the preceding history. Our theoretical and empirical analyses further reveal irreducible future variability and the limitations of fixed Gaussian or convex regions. FutureBound operationalizes this view by using generated trajectories as local probes rather than estimates of normal futures. Fitting and diversity objectives shape a compact yet informative probe set, while selective adversarial refinement progressively sharpens a context-conditioned compatibility boundary. This design enables flexible boundary learning without prescribing its geometry in advance. Quantitative results and boundary visualizations further substantiate the effectiveness of FutureBound and its history-conditioned boundary learning. A current limitation is that the historical and future horizons are fixed across all temporal contexts; adapting the exploration scale to different temporal regimes provides a natural direction for future work.

AI USE STATEMENT

In this work, the conceptual framework and core ideas presented in this paper represent the authors' original contributions. We used generative AI tools solely for language-related assistance, including manuscript translation and linguistic polishing. We have not used generative AI tools for developing the methodology, designing algorithms, conducting experiments, analyzing results, or generating scientific claims, and other generative AI-related tasks are not applicable to this work. We have reviewed all AI-assisted text and ensured that it accurately reflects the authors' original ideas and contributions. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI tools.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the source code and detailed implementation of FutureBound in the supplementary material. The complete algorithm procedure, model configurations, hyper-parameters, training details, and computational resources are provided in Appendix C and Appendix E. All datasets used in our experiments are publicly available, with their statistics summarized in Appendix D. The experimental protocols and baseline settings are described in Section 4.1.

REFERENCES

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3395–3404, 2020.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3):1–33, 2021.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.

Jinyu Cai, Yuan Xie, Glynnis Lim, Yifang Yin, Roger Zimmermann, and See-Kiong Ng. Self-perturbed anomaly-aware graph dynamics for multivariate time-series anomaly detection. *Advances in Neural Information Processing Systems*, 38:84785–84805, 2025.

Chris U. Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. Neural contextual anomaly detection for time series. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2843–2851, 2022.

Zhihao Dai, Ligang He, Shuang-Hua Yang, and Matthew Leeke. Sarad: Spatial association-aware anomaly detection and diagnosis for multivariate time series. *Advances in Neural Information Processing Systems*, 37:48371–48410, 2024.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021.

Cheng Feng and Pengwei Tian. Time series anomaly detection for cyber-physical systems via neural system identification and bayesian filtering. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2858–2867, 2021.

Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2508–2517, 2021.

Alexander Geiger, Dongyu Liu, Sarah Alnegheimish, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Tadgan: Time series anomaly detection using generative adversarial networks. In *Proceedings of the IEEE International Conference on Big Data*, pp. 33–43, 2020.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *Proceedings of the International Conference on Machine Learning*, pp. 3711–3721, 2020.

Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *Proceedings of the 28th International Conference on Artificial Neural Networks*, pp. 703–716, 2019.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: Copula-based outlier detection. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 1118–1123, 2020.

Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3220–3230, 2021.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008.

Zehao Liu, Mengzhou Gao, and Pengfei Jiao. Gcad: Anomaly detection in multivariate time series from the perspective of granger causality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19041–19049, 2025.

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

Kohei Obata, Yasuko Matsubara, and Yasushi Sakurai. Robust and explainable detector of time series anomaly via augmenting multiclass pseudo-anomalies. In *Proceedings of the 30st ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2198–2209, 2025.

Jinju Park and Seokho Kang. Paano: Patch-based representation learning for time-series anomaly detection. In *Proceedings of the 14th International Conference on Learning Representations*, 2026.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 427–438, 2000.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the International Conference on Machine Learning*, pp. 8857–8868, 2021.

Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3009–3017, 2019.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the International Conference on Machine Learning*, pp. 4393–4402, 2018.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. volume 12, 1999.

Lifeng Shen, Zhuocong Li, and James Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems*, 33:13016–13026, 2020.

Lifeng Shen, Liang Peng, Ruiwen Liu, Shuyin Xia, and Yi Liu. Finding time series anomalies using granular-ball vector data description. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25295–25303, 2026.

Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. Memto: Memory-guided transformer for multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57947–57963, 2023.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837, 2019.

Xin Sun, Heng Zhou, Yuhao Wu, and Chao Li. State-derivative-aware neural controlled differential equations for multivariate time series anomaly detection and diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25727–25735, 2026.

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

Herman Wold. *A Study in the Analysis of Stationary Time Series*. PhD thesis, Almqvist & Wiksell, 1938.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

Hongzuo Xu, Yijie Wang, Songlei Jian, Qing Liao, Yongjun Wang, and Guansong Pang. Calibrated one-class classification for unsupervised time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5723–5736, 2024.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *Proceedings of the International Conference on Learning Representations*, 2022.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3033–3045, 2023.

Weiqi Zhang, Chen Zhang, and Fugee Tsung. Grelen: Multivariate time series anomaly detection from the perspective of graph relational learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 734, pp. 2390–2397, 2022.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *IEEE International Conference on Data Mining*, pp. 841–850, 2020.

A Proof of Proposition 1

Proof. Since $\boldsymbol{\mu}_t = \mathbb{E}_{\mathrm{N}}[\mathbf{F}_t|\mathbf{H}_t]$, its first future state is $\mathbb{E}_{\mathrm{N}}[\mathbf{x}_t|\mathbf{H}_t]$. Using $\mathbf{x}_t = g(\mathbf{H}_t) + \boldsymbol{\varepsilon}_t$ and $\mathbb{E}_{\mathrm{N}}[\boldsymbol{\varepsilon}_t|\mathbf{H}_t] = \mathbf{0}$, we have

$$\mathbb{E}_{\mathrm{N}}[\mathbf{x}_t|\mathbf{H}_t] = g(\mathbf{H}_t), \tag{13}$$

and therefore

$$\mathbf{x}_t - \mathbb{E}_{\mathrm{N}}[\mathbf{x}_t|\mathbf{H}_t] = \boldsymbol{\varepsilon}_t. \tag{14}$$

Since the squared residual of the complete future trajectory includes the contribution from its first future state, we have

$$V_t = \mathbb{E}_{\mathrm{N}}\left[\|\mathbf{F}_t - \boldsymbol{\mu}_t\|_F^2 | \mathbf{H}_t\right] \tag{15}$$

$$\geq \mathbb{E}_{\mathrm{N}}\left[\|\mathbf{x}_t - \mathbb{E}_{\mathrm{N}}[\mathbf{x}_t|\mathbf{H}_t]\|_2^2 \,|\mathbf{H}_t\right] \tag{16}$$

$$= \mathbb{E}_{\mathrm{N}}\left[\|\boldsymbol{\varepsilon}_t\|_2^2 | \mathbf{H}_t\right]. \tag{17}$$

Taking expectation over the historical context yields

$$\mathbb{E}_{\mathrm{N}}[V_t] \geq \mathbb{E}_{\mathrm{N}}\left[\|\boldsymbol{\varepsilon}_t\|_2^2\right] > 0. \tag{18}$$

For any deterministic predictor $f(\mathbf{H}_t)$, the conditional expectation $\boldsymbol{\mu}_t = \mathbb{E}_{\mathrm{N}}[\mathbf{F}_t|\mathbf{H}_t]$ minimizes the expected squared prediction error. Hence,

$$\inf_f \mathbb{E}_{\mathrm{N}}\left[\|\mathbf{F}_t - f(\mathbf{H}_t)\|_F^2\right] = \mathbb{E}_{\mathrm{N}}\left[\|\mathbf{F}_t - \boldsymbol{\mu}_t\|_F^2\right] \tag{19}$$

$$= \mathbb{E}_{\mathrm{N}}[V_t] > 0. \tag{20}$$

This completes the proof. □

B Discussion on Asymmetric Probe Selection

The proposed FutureBound deliberately uses different probe subsets for discriminator and generator optimization. The discriminator is trained on a relatively broader set of nearby probes, $\mathcal{I}_t^{\mathcal{D}}$, selected solely according to the trajectory discrepancy $d_t^{(m)}$. This provides sufficient local coverage around the observed normal continuation and allows the discriminator to establish a stable context–future distinction without being distracted by distant trajectories. In contrast, the generator updates only a small subset $\mathcal{I}_t^{\mathcal{G}}$ selected from different residual orientations. This selective refinement preserves the exploratory role of the remaining probes while making a few nearby candidates progressively more challenging to the discriminator. Therefore, $\mathcal{I}_t^{\mathcal{D}}$ and $\mathcal{I}_t^{\mathcal{G}}$ serve complementary purposes: the former supports broad local discrimination, whereas the latter promotes targeted boundary refinement.

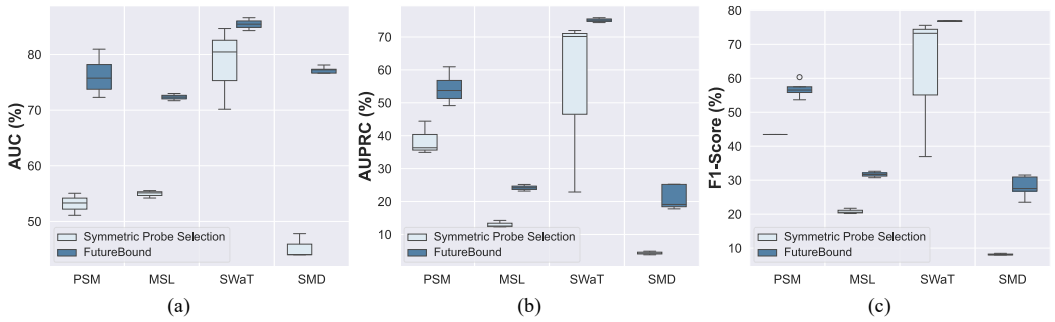


Figure 7: Comparison between asymmetric and symmetric probe selection on four datasets.

Here we further introduce a variant termed **Symmetric Probe Selection**, which applies the same orientation-aware probe selection to both discriminator and generator optimization while keeping the same probe budget, allowing us to assess the effect of their distinct selection criteria. As shown in Fig. 7, enforcing a shared selection strategy causes substantial degradation on PSM, MSL, and SMD, while also increasing the variance on SWaT, particularly for AUPRC and F1-Score. This suggests that the two adversarial steps benefit from different probe-selection principles. The discriminator is better served by selecting the nearest probes without directional constraints, thereby preserving the natural local neighborhood around the observed continuation, whereas the generator benefits from explicitly retaining different residual orientations to avoid concentrating adversarial refinement on only one side. Applying the orientation constraint to both steps alters the discriminator's local probe distribution and leads to less stable and less effective boundary learning.

C Detailed Algorithm Procedures

The complete optimization and inference procedure of FutureBound is summarized in Algorithm 1. The algorithm details how historical contexts are encoded, how multiple future probes are constructed and refined through adversarial learning, and how the learned compatibility scores are finally converted into timestamp-level anomaly scores. The separate discriminator and generator steps further clarify the alternating optimization used to learn the implicit future boundary.

Algorithm 1 Detailed Procedures of the proposed FutureBound.

Require: Multivariate time series $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, historical length L, future horizon H, number of future candidates M, discriminator selection size $K_{\mathcal{D}}$, generator selection size $K_{\mathcal{G}}$, transition weight λ_{Δ}, diversity scale σ_{div}, loss weights α and β, timestamp aggregation size k.

Ensure: Timestamp-level anomaly scores $\mathbf{S} = \{s_\tau\}$.

Construct historical-future pairs $(\mathbf{H}_t, \mathbf{F}_t)$ from the training sequence;
Initialize the TCN encoder E, multi-future generator $\mathcal{G}_\theta$, and context-conditioned discriminator $\mathcal{D}_\psi$;
// Boundary-Oriented Adversarial Training
while not converged **do**
  Sample a batch of normal historical–future pairs from $\mathcal{S}_{\text{train}}$;
  for each pair $(\mathbf{H}_t, \mathbf{F}_t)$ in the batch **do**
    Obtain the temporal context representation $\mathbf{h}_t = E(\mathbf{H}_t)$ following Eq. (2);
    Generate M context-dependent future trajectories $\{\hat{\mathbf{F}}_t^{(m)}\}_{m=1}^{M}$ with Eq. (3);
    Calculate the trajectory discrepancy $d_t^{(m)}$ for each candidate using Eq. (4);
    Choose the $K_{\mathcal{D}}$ candidates with the smallest $d_t^{(m)}$ and form $\mathcal{I}_t^{\mathcal{D}}$;
  end for
  Calculate the fitting loss $\mathcal{L}_{\text{fit}}$ and diversity loss $\mathcal{L}_{\text{div}}$ using Eqs. (5) and (6);
  Calculate $\mathcal{L}_{\text{dis}}$ over the observed futures and selected nearby probes using Eq. (7), and optimize the discriminator parameters ψ;
  for each pair $(\mathbf{H}_t, \mathbf{F}_t)$ in the batch **do**
    Calculate the signed mean residual $c_t^{(m)}$ using Eq. (8), and divide the candidates into $\mathcal{U}_t$ and $\mathcal{L}_t$;
    Select the $K_{\mathcal{G}}$ nearest candidates from each residual side according to $d_t^{(m)}$, and form the adversarial probe set $\mathcal{I}_t^{\mathcal{G}}$;
    Fill any missing positions with the nearest unused candidates when either residual side contains fewer than $K_{\mathcal{G}}$ candidates;
  end for
  Fix the discriminator and calculate the adversarial loss $\mathcal{L}_{\text{adv}}$ using Eq. (9);
  Assemble the generator objective $\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\text{adv}} + \alpha\mathcal{L}_{\text{fit}} + \beta\mathcal{L}_{\text{div}}$ following Eq. (10);
  Back-propagate $\mathcal{L}_{\mathcal{G}}$ and optimize the generator and temporal encoder;
end while
// Timestamp-Level Anomaly Scoring
for each valid test anchor t **do**
  Construct $\mathbf{H}_t$ and its observed future trajectory $\mathbf{F}_t$, and obtain $\mathbf{h}_t = E(\mathbf{H}_t)$;
  Calculate the trajectory-level anomaly score $a_t = -\mathcal{D}_\psi(\mathbf{h}_t, \mathbf{F}_t)$ using Eq. (11);
end for
for each timestamp τ **do**
  Collect the covering trajectory scores $\{a_t : t \in \mathcal{W}(\tau)\}$;
  Retain the largest $k_\tau = \min(k, |\mathcal{W}(\tau)|)$ scores and obtain s_τ using Eq. (12);
end for
return The anomaly scores $\mathbf{S} = \{s_\tau\}$.

C.1 Computational Complexity Analysis

Assume that the training set contains N historical-future pairs, where each historical context has length L, each future trajectory has horizon H, and the number of variables is D. Let d_h denote the hidden dimension of the temporal encoder, L_{Enc} the number of TCN layers, Q the convolution

kernel size, and M the number of generated future trajectories. The computational complexity of FutureBound is analyzed as follows:

- The temporal context encoder adopts an L_{Enc}-layer TCN. The first layer transforms the D input variables into the hidden space with complexity $\mathcal{O}(LQDd_h)$, while the remaining layers require $\mathcal{O}(LQd_h^2)$ each. Therefore, the overall encoding complexity is $\mathcal{O}(LQ(Dd_h + (L_{\text{Enc}} - 1)d_h^2))$ for each historical context.
- The multi-future generator produces M trajectories, each containing $H \times D$ values. The output projection therefore requires $\mathcal{O}(MHDd_h)$ operations. Computing the state and transition discrepancies between the generated trajectories and the observed future further requires $\mathcal{O}(MHD)$.
- The diversity regularizer evaluates the pairwise discrepancy among the M generated trajectories. Since each pair involves HD elements, its computational complexity is $\mathcal{O}(M^2HD)$.
- For discriminator optimization, the $K_{\mathcal{D}}$ nearest trajectories are selected from the M generated candidates. The selection requires at most $\mathcal{O}(M \log M)$ operations. Assuming that scoring one context–future pair requires $\mathcal{O}(HDd_h)$ operations, evaluating the observed future and the selected probes requires $\mathcal{O}(K_{\mathcal{D}}HDd_h)$.
- For adversarial generator optimization, calculating the signed residuals of all candidates requires $\mathcal{O}(MHD)$. The candidates are further ranked within the two residual subsets with complexity $\mathcal{O}(M \log M)$. Since at most $2K_{\mathcal{G}}$ probes are used for adversarial optimization, the corresponding discriminator evaluation requires $\mathcal{O}(K_{\mathcal{G}}HDd_h)$.

Since M, $K_{\mathcal{D}}$, and $K_{\mathcal{G}}$ are small fixed hyper-parameters, the additional future-space exploration grows linearly with the future horizon H and the number of variables D, except for the quadratic dependence on M introduced by the diversity regularizer.

During inference, the generator and the pairwise diversity computation are no longer required. For each test trajectory, FutureBound only performs temporal context encoding and context–future discrimination. Hence, for N_{test} test windows, the inference complexity is $\mathcal{O}\left(N_{\text{test}}\left[LQ(Dd_h + (L_{\text{Enc}} - 1)d_h^2) + HDd_h\right]\right)$. The subsequent timestamp-level top-k aggregation considers at most H overlapping trajectory scores for each timestamp, resulting in an additional complexity of $\mathcal{O}(T_{\text{test}}H \log k)$.

C.2 Running Time Comparison

To further complement our efficiency analysis, we compare the training time of FutureBound with representative baselines on PSM and MSL in Fig. 8. For a fair comparison of per-epoch training efficiency, all methods are trained for 30 epochs under the same experimental environment, and the cumulative training time over these 30 epochs is recorded. The results reveal substantial differences in computational cost across different modeling paradigms. For example, GCAD additionally performs gradient-based causal discovery and graph construction, while SPAGD combines self-perturbed sequence generation, dynamic graph construction, and spatio-temporal discrimination. USAD adopts a dual-autoencoder architecture with adversarial reconstruction, resulting in a different optimization procedure from conventional reconstruction models. Despite incorporating multi-future generation and adversarial boundary learning, FutureBound remains among the most efficient methods on both datasets. This efficiency benefits from its compact design: candidate trajectories are generated jointly from a shared historical representation, and adversarial refinement is restricted to only a small subset of selected probes. Consequently, FutureBound introduces additional modeling flexibility without incurring excessive training overhead.

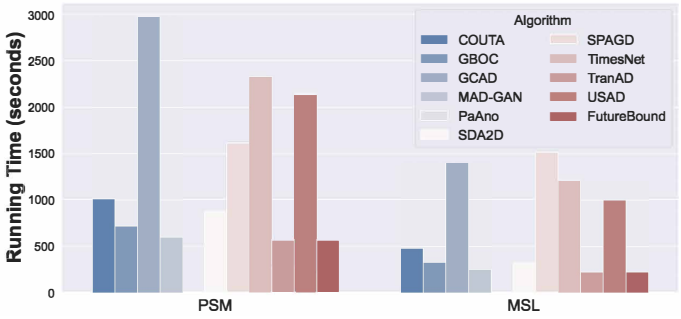


Figure 8: Running time comparison between several SOTAs and the proposed model on PSM and MSL datasets.

D DATASETS

The detailed descriptions of four multivariate time-series datasets, including SWaT, PSM, MSL, SMD, are displayed in Table 2.

Table 2: Statistical descriptions of used time-series datasets.

Dataset	Domain	# Variables	# Train Samples	# Test Samples	Anomaly Rate (%)
SWaT	Water Treatment ICS	51	496,800	449,919	12.14
PSM	Application Server Metrics	25	132,481	87,841	27.76
MSL	Spacecraft Telemetry	55	58,317	73,729	10.53
SMD	Server Machine Metrics	38	708,405	708,420	4.16

E DETAILED EXPERIMENTAL SETTINGS

- **Data Splits:** We conduct all experiments on SWaT, PSM, SMD, and MSL using their official train/test splits. Only normal training data are used to learn the history-conditioned future space, and each variable is standardized using statistics estimated from the training split. At testing stage, the first L observations are used as a warm-up prefix to initialize the historical context and are excluded from evaluation.
- **Model Architecture:** The temporal context encoder is implemented with a four-level TCN with hidden dimension $d_h = 64$, kernel size 3, dilation factors $\{1, 2, 4, 8\}$. Given the context representation $\mathbf{h}_t$, the context-conditioned generator produces $M = 4$ candidate future trajectories. Each candidate is a complete multivariate trajectory in $\mathbb{R}^{H\times D}$. The context-conditioned discriminator $\mathcal{D}_\psi(\mathbf{h}_t, \mathbf{F})$ is implemented as an MLP over the concatenation of $\mathbf{h}_t$ and the flattened future trajectory, with hidden dimension 64. Spectral normalization is applied to all linear layers to control the scale of the discriminator function.
- **Hyper-parameter Settings:** We set the historical context length to $L = 100$ and the future horizon to $H = 16$. Training windows are extracted with stride 10, whereas test windows use stride 1 to obtain dense timestamp-level anomaly scores. The trajectory transition weight is set to $\lambda_\Delta = 1$. The diversity regularization weight and diversity scale are set to $\beta = 0.01$ and $\sigma_{\text{div}} = 1$, respectively. The fitting loss weight α is set to 1. For local probe selection, we use $K_\mathcal{D} = 2$ nearby probes for discriminator optimization and $K_\mathcal{G} = 1$ probe from each residual orientation for generator refinement.
- **Training Details:** FutureBound is trained for 100 epochs with batch size 128. AdamW (Loshchilov & Hutter, 2019) is used for optimization, with learning rates of 10^{-3} for the temporal encoder and generator and 10^{-4} for the discriminator. The discriminator and generator are updated alternately with one discriminator update per generator update, and the gradient norm is clipped to 1.0.
- **Inference and Evaluation:** Overlapping trajectory scores are aggregated into timestamp-level scores using top-k pooling with $k = 3$. The evaluation batch size is set to 512. Please note that each trajectory-level score is produced once its corresponding H-step future window has been fully observed, resulting in a maximum detection latency of $H - 1$ timestamps.
- **Baseline Settings:** For all baseline methods, we use publicly available implementations whenever possible and follow the hyper-parameter settings recommended by the original papers or their official repositories[12]. Each experiment is repeated ten times with independently generated random seeds.
- **Computational Resources:** All experiments are conducted on a server equipped with an Intel Xeon Gold 5318Y CPU and an NVIDIA A800 PCIe GPU with 80GB memory.

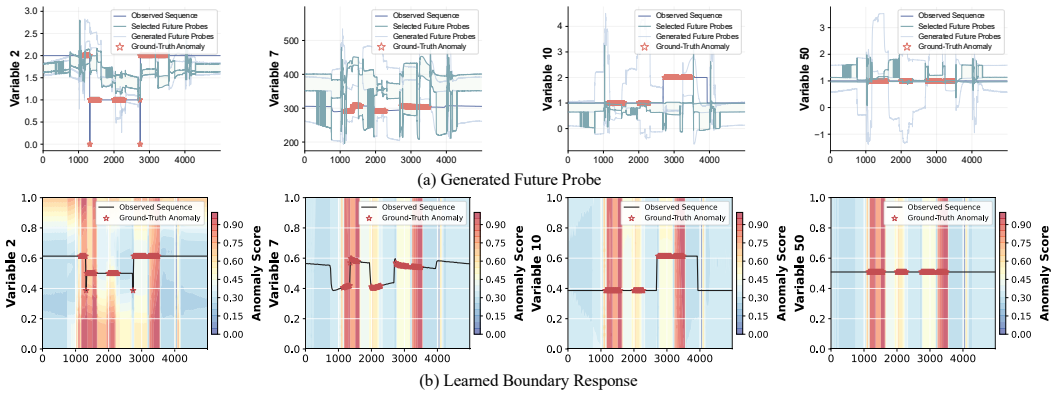


Figure 9: Decision boundary visualizations of FutureBound on representative SWaT variables. (a) Generated and selected future probes on testing data; (b) Learned boundary responses corresponding to the observed testing sequences.

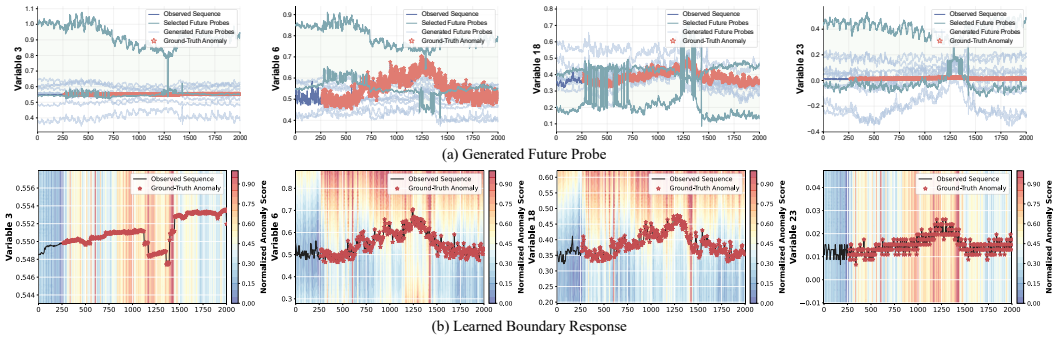


Figure 10: Decision boundary visualizations of FutureBound on representative PSM variables. (a) Generated and selected future probes on testing data; (b) Learned boundary responses corresponding to the observed testing sequences.

F Supplementary Visualization Results

Additional Decision Boundary Visualizations. We further provide decision-boundary visualizations on additional SWaT variables and representative PSM variables in Figs. 9 and 10. The SWaT results complement the partial visualization in the main text and show consistent behavior across variables with different operating ranges and temporal patterns. The generated probes remain concentrated around normal evolution while expanding around regime transitions, and the learned boundary responses become markedly stronger over anomalous intervals. This consistency across multiple variables indicates that the observed boundary behavior is not specific to the examples shown in the main text. On PSM, several variables exhibit smoother and more gradual temporal changes, making anomalous behavior less distinguishable from the raw observations alone. The generated probes generally follow the prevailing operating levels, while noticeable departures appear around regime changes, particularly for Variables 6 and 18. The discriminator responses make these deviations more evident: anomalous intervals are largely aligned with warm response regions, whereas long stable periods remain predominantly blue. Variable 23 is particularly illustrative, where relatively small changes in absolute magnitude still produce clear response variations, indicating that the learned boundary captures deviations from the local temporal context rather than

[1]https://github.com/thedatumorg/TSB-AD/tree/main
[2]https://github.com/NAThompson/DeepOD/tree/main

relying solely on value magnitude. Together, the visualizations demonstrate that FutureBound produces consistent boundary responses under both abrupt and gradual temporal variations.

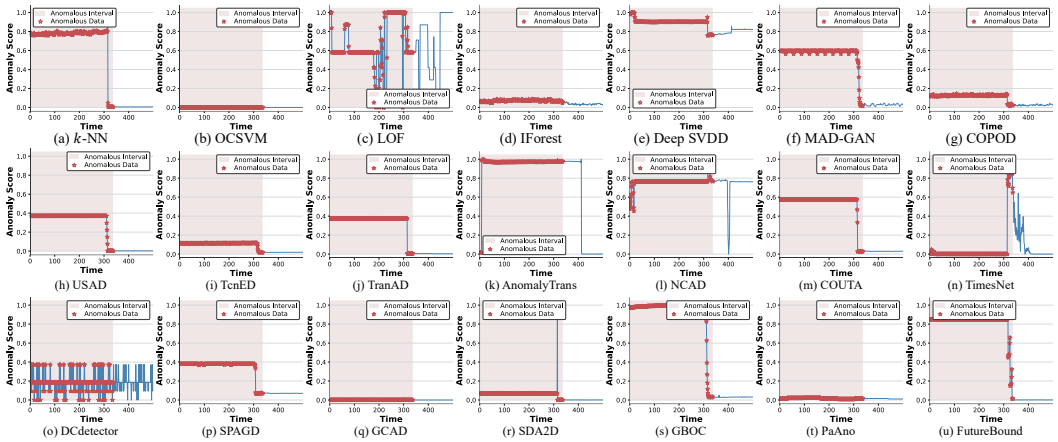


Figure 11: Complete anomaly score visualization comparison on a 500-point SWaT test segment. Blue curves represent timestamp-level anomaly scores, red shading indicates ground-truth anomalous intervals, and red stars denote anomalous timestamps.

Full Detector Behavior Comparison. We provide the complete timestamp-level anomaly score visualization of all compared methods on the same 500-point SWaT test segment in Fig. 11. The full comparison reveals diverse response behaviors across detectors. Some methods, such as LOF and DCdetector, generate fragmented activations within the anomalous interval, while others, including Deep SVDD and SPAGD, show limited sensitivity to the transition between anomalous and normal regimes. MAD-GAN, GBOC, and TimesNet provide relatively clearer responses but still exhibit either residual activations or unstable score fluctuations. In contrast, FutureBound consistently assigns high anomaly scores throughout the anomalous interval and rapidly returns to a low response after the system resumes normal evolution. This stable transition demonstrates the benefit of modeling anomalies through history-conditioned future compatibility, where deviations are identified according to whether the observed future remains within the learned normal evolution boundary.

G Supplementary Parameter Sensitivity

This section supplements the main-text analysis with the complete parameter sensitivity results on SMD and PSM across AUC, AUPRC, and F1-Score, as shown in Figs. 12 and 13. The complete results reveal additional observations: **(1)** For M and H, the results reveal a clear coupling between exploration breadth and temporal extent. Increasing M provides denser coverage of the local future space, but its benefit gradually saturates when the generated probes become redundant. Increasing H exposes longer-term temporal evolution, yet an excessively long horizon also enlarges the trajectory space, making local fitting and boundary refinement more difficult. Consequently, neither additional probes nor a longer horizon can compensate indefinitely for the other. **(2)** The two loss weights exhibit clearly asymmetric sensitivity. Performance is considerably more sensitive to α than to β: excessively large fitting weights cause substantial degradation, especially on PSM, while β can vary over a relatively wider range with much smaller changes when α remains moderate. This suggests that maintaining an appropriate degree of local anchoring is more critical than precisely tuning the diversity strength.

H Supplementary Ablation Results

Fig. 14 presents the complete ablation results in terms of AUC, AUPRC, and F1-Score. In addition to the AUC and AUPRC results reported in the main text, the F1-Score results exhibit consistent trends

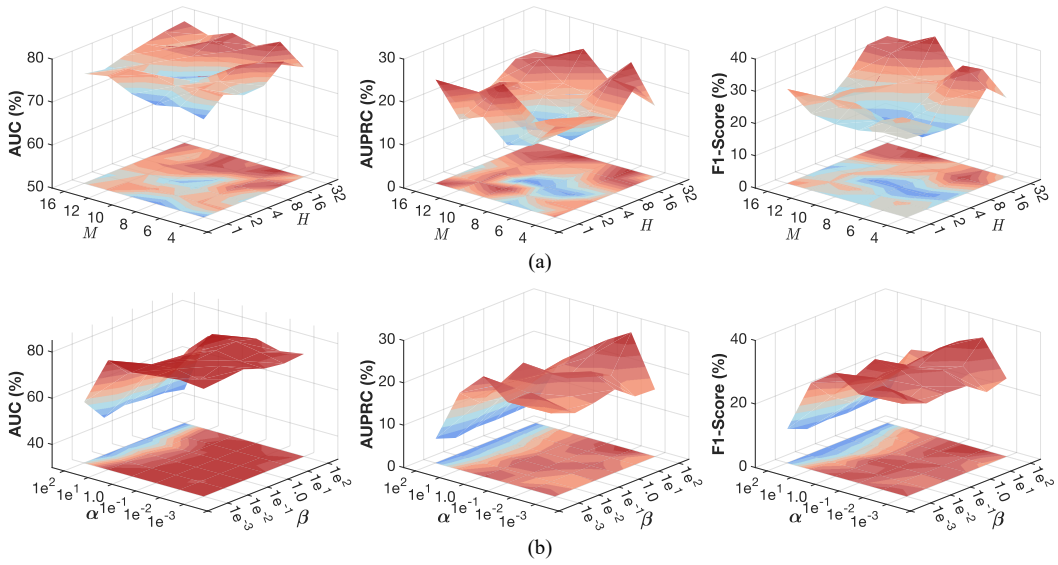


Figure 12: Complete parameter sensitivity results on SMD across AUC, AUPRC, and F1-Score. (a) Sensitivity to the number of generated trajectories M and future horizon H. (b) Sensitivity to the weights α and β of the fitting and diversity losses.

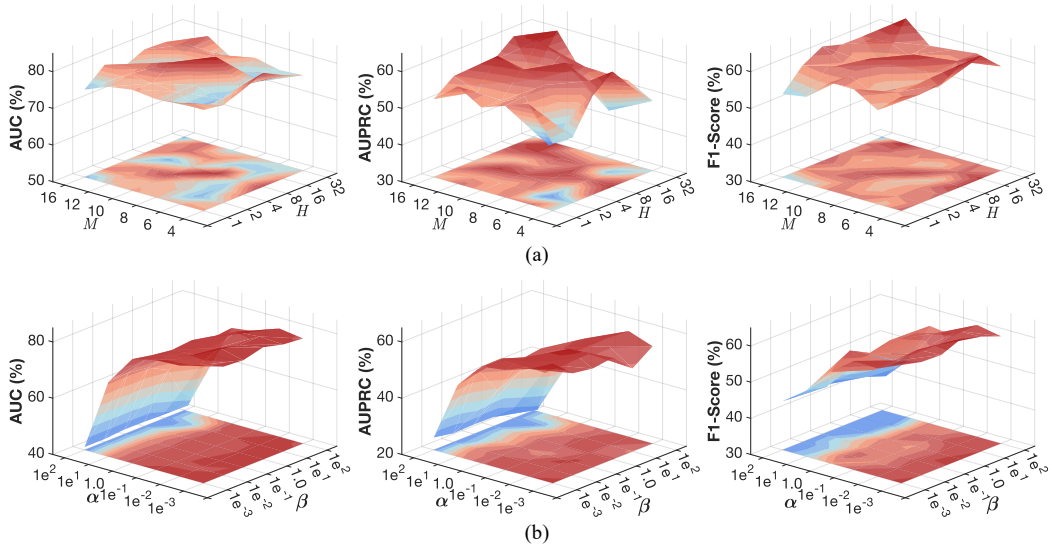


Figure 13: Parameter sensitivity results on PSM across AUC, AUPRC, and F1-Score. (a) Sensitivity to the number of generated trajectories M and future horizon H. (b) Sensitivity to the weights α and β of the fitting and diversity losses.

across the four datasets. In particular, the complete model generally outperforms its variants, while removing the discriminator, diversity regularization, or fitting constraint leads to varying degrees of degradation. Gaussian Perturbation-GAN also remains less consistent across datasets, further supporting the use of learned context-conditioned probes for boundary refinement.

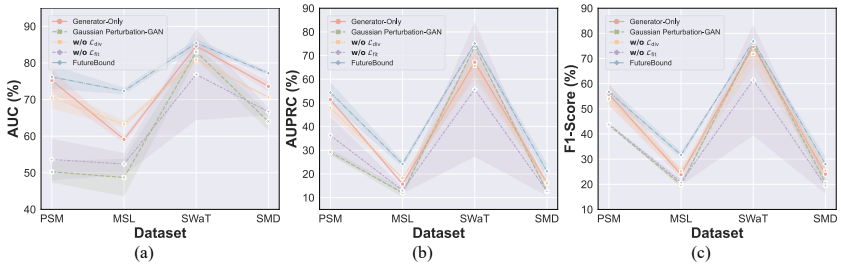


Figure 14: Complete ablation results of FutureBound across four datasets in terms of AUC, AUPRC, and F1-Score. Shaded regions indicate standard deviations over 10 runs.

I Related Work

I.1 Prediction, Reconstruction, and Probabilistic Modeling

A major line of time-series anomaly detection characterizes normal temporal behavior through prediction or reconstruction fidelity. OmniAnomaly (Su et al., 2019) models normal multivariate dynamics with stochastic latent variables and evaluates anomalies through reconstruction probabilities, while GDN (Deng & Hooi, 2021) explicitly captures inter-variable dependencies for future prediction. InterFusion (Li et al., 2021) further models temporal and cross-variable dependencies through hierarchical stochastic representations, and transformer-based methods such as TranAD (Tuli et al., 2022) and MEMTO (Song et al., 2023) improve reconstruction by incorporating attention and memory mechanisms. Beyond anomaly detection, probabilistic forecasting methods explicitly model future uncertainty rather than collapsing temporal evolution into a single prediction. For example, TimeGrad (Rasul et al., 2021) learns a history-conditioned predictive distribution through autoregressive diffusion, from which multiple possible future trajectories can be sampled. Flexible probabilistic models can in principle capture multi-modal distributions with complex, non-convex high-density regions and are therefore not restricted to the geometry induced by a single Gaussian model. Their primary objective, however, remains conditional predictive distribution estimation; applying such models to anomaly detection still requires an additional criterion to translate predictive uncertainty into anomaly evidence. FutureBound instead bypasses this intermediate step and directly learns a history-conditioned compatibility boundary tailored to anomaly detection.

I.2 Representation and Dependency Discrepancy

Another family of methods seeks more discriminative criteria in the learned temporal representation or dependency structure. Anomaly Transformer (Xu et al., 2022) identifies anomalies through discrepancies between temporal associations, while DCdetector (Yang et al., 2023) learns discriminative multi-scale representations through contrastive dual-attention modeling. For multivariate series, GReLeN (Zhang et al., 2022) detects abnormalities through learned inter-sensor relational structures, and SARAD (Dai et al., 2024) models changing spatial associations between variables and detects deviations in the resulting association space. These approaches move beyond raw prediction or reconstruction errors by exploiting irregularities in representations, associations, or variable dependencies, but the detection criterion is still defined through deviations from learned normal structures.

I.3 One-Class and Boundary-based Learning

One-class anomaly detection directly characterizes the region occupied by normal data. Deep SVDD (Ruff et al., 2018) learns a compact hyperspherical representation of normal samples, while DROCC (Goyal et al., 2020) improves one-class discrimination by generating adversarial points around normal observations. For time series, THOC (Shen et al., 2020) models temporal dynamics through multiple hierarchical hyperspheres, and COUTA (Xu et al., 2024) calibrates one-class learning against contamination and generated native anomalies. Recent geometric approaches such as GBOC (Shen et al., 2026) further represent normal temporal patterns using multiple granular-ball structures. These methods provide an explicit boundary-oriented view of normality, but their

decision regions are generally constructed in a learned representation space or within a prescribed geometric family. In contrast, FutureBound learns a boundary over future trajectories conditioned on the preceding history, allowing the effective normal region to change across temporal contexts without fixing its geometry in advance.

I.4 Generative and Synthetic-Anomaly-based Learning

Generation has also commonly been used to provide additional supervision for anomaly detection. Early adversarial approaches such as MAD-GAN (Li et al., 2019) use generative modeling to learn normal sequence distributions. NCAD (Carmona et al., 2022) instead injects generic synthetic anomalies and learns contextual discrimination between normal and perturbed windows. More recent approaches further strengthen this paradigm: RedLamp (Obata et al., 2025) constructs multiple classes of pseudo-anomalies to improve robustness under contaminated training data, while SPAGD (Cai et al., 2025) generates self-perturbed multivariate sequences and exploits them to learn anomaly-aware graph dynamics. FutureBound assigns a fundamentally different role to generated trajectories. They are neither samples intended to approximate the normal distribution nor synthetic anomalies with assumed abnormal labels; instead, they serve as locally anchored probes whose interaction with observed normal continuations provides information for refining the surrounding admissible boundary.