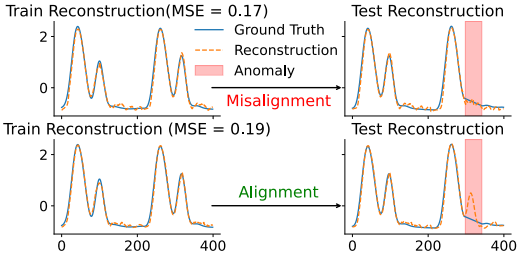
AnoMamba: Aligning Reconstruction with Time Series Anomaly Detection via Selective Global Dependency Modeling

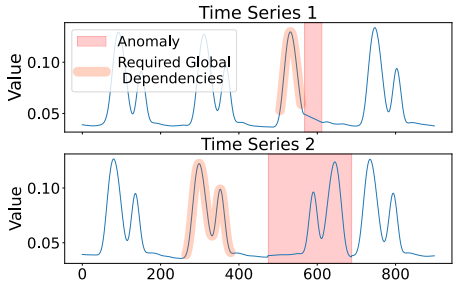
Junqi Chen[1,2] , **Xu Tan**[1] , **Jie Chen**[1,2*] and **Susanto Rahardja**[3*]
[1]Northwestern Polytechnical University
[2]Research & Development Institute of Northwestern Polytechnical University in Shenzhen
[3]Zhejiang University
{jqchen, xutan, dr.jie.chen, susantorahardja}@ieee.org

Abstract

Reconstruction-based frameworks are widely adopted in Time Series Anomaly Detection (TSAD), assuming that models reconstruct normal behavior well but yield larger errors on anomalies. However, in unsupervised TSAD, minimizing reconstruction loss alone often breaks this assumption. Models tend to overfit local patterns for trivial reconstruction and fail to capture the global dependencies that characterize normal behavior. Consequently, anomalies that violate global dependencies can also be reconstructed well, leading to a misalignment between reconstruction and detection. To address this challenge, we propose AnoMamba, a novel TSAD framework that aligns reconstruction with anomaly detection by enhancing global dependency modeling. Specifically, AnoMamba employs patch embedding to reduce local redundancy and introduces a Mamba variant with global step-size reweighting to select meaningful global dependencies. The reweighting process is guided by multi-scale long-tail priors, which adaptively balance global and local dependencies to mitigate overfitting in the unsupervised setting. Extensive experiments on both univariate and multivariate benchmarks demonstrate that AnoMamba consistently outperforms state-of-the-art methods in both accuracy and efficiency, while offering interpretability through its hidden attention map.



(a) Misalignment between reconstruction and anomaly detection. Top: the detector overfits to local patterns and fails to learn normal behavior, leading to misalignment. Bottom: expected alignment, where reconstruction fits normal behavior but not anomalies. MSE denotes mean square error.



(b) Variable and ambiguous boundaries between global and local dependencies: The global dependencies in Time Series 1 may become local dependencies in Time Series 2.

Figure 1: Challenges in reconstruction-based TSAD.

1 Introduction

Time Series Anomaly Detection (TSAD) has attracted increasing attention due to its critical role in a wide range of real-world applications, including medical diagnosis, fault detection, and instruction detection [Zamanzadeh Darban *et al.*, 2024]. Among various approaches, reconstruction-based frameworks are widely-adopted in TSAD, as they can detect anomalies by measuring reconstruction errors under the core assumption that only normal patterns can be accurately reconstructed [Schmidl *et al.*, 2022]. Despite the integration of powerful sequence models over the past decade, reconstruction-based methods remain fundamentally susceptible to overfitting [Boniol *et al.*, 2024]. In particular, an overfitted detector may reconstruct anomalies with high accuracy, thereby violating the foundational assumption of reconstruction-based TSAD.

This overfitting issue stems from a fundamental misalignment between the reconstruction objective and the anomaly detection objective. In TSAD, the goal of reconstruction is not merely to minimize reconstruction error, but to learn a reliable normal behavior. However, in the absence of labeled anomalies, optimizing reconstruction loss alone en-

*Corresponding Authors

courages models to overfit local dependencies, which are easier to reconstruct [Xu *et al.*, 2021]. As a result, the learned model fails to fully characterize normal behavior, leading to the unexpected reconstruction of anomalies. As shown in Fig. 1 (a), this local overfitting issue directly undermines the anomaly detection objective and raises a core challenge for reconstruction-based TSAD: **Challenge1: The misalignment between reconstruction and anomaly detection hinders the normal behavior learning.**

A natural solution to alleviate this misalignment is to enforce the learning of global dependencies. By modeling global dependencies, the model is encouraged to learn more complete normal behavior beyond trivial local patterns. Existing methods lie in two perspectives. Feature-level methods construct multi-scale inputs, such as multi-resolution views [Wang *et al.*, 2024b; Zhong *et al.*, 2025] and decomposition-based components [Wu *et al.*, 2021; Chen *et al.*, 2024], to encourage detectors learn global dependencies residing in coarse resolutions or low-frequency components. However, the feature construction is not trained end-to-end (E2E) with the detector, making performance highly sensitive to hyperparameters. In contrast, objective-level methods integrate regularization to E2E guide the detector to capture global dependency. Typical methods use attention priors [Xu *et al.*, 2021] or global windows [Chen *et al.*, 2025; Peng *et al.*, 2025] to constrain the detector's attentions to predefined global regions. Despite being E2E, these designs still fix in advance where the model should attend, imposing overly strong assumptions on the global regions. In practice, the required global dependencies vary across time and context, as illustrated in Fig. 1 (b). This leads to the second challenge: **Challenge2: It is difficult to define global dependencies for normal behavior in advance.**

To tackle the above challenges, we explore alternative solutions from the recently proposed Mamba model [Gu and Dao, 2024]. A key characteristic of it is its input-dependent step-size, which controls the contribution of historical information and enables the selectively modeling of global dependencies. However, existing applications of Mamba to time series are predominantly in supervised settings [Xu *et al.*, 2024; Karadag *et al.*, 2025], where the step-size is guided by labels. In unsupervised TSAD, this selective mechanism cannot be effectively exploited, leaving the challenges unsolved.

To fill this gap, we propose Global Step-size Reweighting Mamba (GSRMamba), a Mamba variant tailored for unsupervised TSAD. The design centers on controlling step-sizes via global context and multi-scale priors. First, we introduce a global step-size reweighting model, which modulates the step-size with global contextual information. Second, we develop a multi-scale prior regularization strategy to guide the learning of this reweighting process. Unlike previous attention priors that directly impose importance on predefined positions, our step-size priors only balance the importance of global and local dependencies without making positional assumptions. We further associate the reweighted step-sizes with a hidden attention map [Ali *et al.*, 2025] to enhance the interpretability of normal behavior learning. Finally, we incorporate patch embedding into GSRMamba to reduce local dependency redundancy. These components are integrated into a unified reconstruction-based TSAD framework, termed AnoMamba, which enables capturing variable global dependencies and realigns reconstruction with anomaly detection.

The contributions of this paper are summarized as follows:

- We propose AnoMamba, a novel reconstruction-based anomaly detector that enhances global dependencies modeling to align reconstruction with TSAD.
- To address the challenges of variable global dependencies modeling, we introduce GSRMamba, which leverages step-size control and multi-scale prior guidance to fully exploit Mamba's selective capability for precise global information selection.
- We connect the learned step-size with a hidden attention map to enhance the interpretability of normal behavior learning.
- Extensive experiments on both univariate and multivariate datasets demonstrate the superior detection performance and efficiency of AnoMamba compared to state-of-the-art (SOTA) methods.

2 Related Works

In this section, we review reconstruction-based TSAD methods and recent Mamba-based approaches for TSAD.

2.1 Reconstruction-based TSAD

Reconstruction-based TSAD aims to learn normal behavior by training a model to reconstruct the input series. Early reconstruction ideas can be traced back to PCA [Hoffmann, 2007] and autoencoder-based methods [An and Cho, 2015], where normal patterns are captured by principal components or low-dimensional representations. However, these methods ignore temporal structure and thus struggle with time series contained strong contextual dependencies. To model temporal dependencies, reconstruction is typically performed on windowed sequences, and various deep sequential architectures have been explored, including Multi-Layer Perceptron (MLP)-based [Zeng *et al.*, 2023], Convolutional Neural Network (CNN)-based [Wu *et al.*, 2023], Recurrent Neural Network (RNN)-based [Su *et al.*, 2019], and Transformer-based [Xu *et al.*, 2021] models. MLPs and CNNs are computationally efficient, whereas RNNs and Transformers are more suitable for long-range dependencies. More recently, State Space Models (SSM), such as Mamba [Gu and Dao, 2024], have been proposed to strike a balance between efficiency and modeling power.

Despite these advances, reconstruction-based TSAD remains vulnerable to overfitting to local patterns, due to the misalignment between reconstruction and anomaly detection in unsupervised settings. Existing methods attempt to alleviate this by emphasizing global dependencies, but they often rely on strong assumptions such as fixed positional priors [Xu *et al.*, 2021] or windowing schemes [Chen *et al.*, 2025]. In this work, we instead develop a Mamba variant that captures meaningful global dependencies required for normal behavior learning. Our design enables the model to adaptively focus on different global ranges, thereby mitigating overfitting and improving anomaly detection performance.

2.2 Mamba for TSAD

Mamba [Gu and Dao, 2024] is a recently proposed variant of SSMs that overcomes the time-invariance limitation present in earlier SSMs. Its core innovation lies in the introduction of the selective SSM (S6), which incorporates time-dependent parameters into the SSM process. Building on this, recent works such as SST [Xu *et al.*, 2024] and DMamba [Chen *et al.*, 2024] have replaced Transformers with Mamba to better capture global dependencies in time series. Among the time-dependent parameters in Mamba, the step-size plays a critical role in controlling the contribution of history information, enabling global information selection. MS-Mamba [Karadag *et al.*, 2025] further scales the step-size with a learned scalar to modulate this effect. However, these methods still rely on supervised training and lack fine-grained control at the individual time-step level. When trained in an unsupervised setting, Mamba-based models still tend to overfit local patterns and cannot fully exploit their inherent selective capability.

In this paper, we introduce a global step-size reweighting model that adaptively adjusts step-sizes across time, enabling finer control of Mamba's selective behavior for global dependency modeling. We further propose a multi-scale prior regularization strategy to guide this reweighting process in the unsupervised setting, effectively alleviating local overfitting. In addition, by linking the learned step-sizes to hidden attention maps, our method offers interpretable insights into the learned normal behaviors, improving the transparency of the anomaly detection process.

3 Preliminaries

In this section, we define the TSAD problem and present the Mamba formulation together with its hidden attention map.

3.1 The Task Definition of TSAD

Given a multivariate time series $\mathbf{x}_t \in \mathbb{R}^D$ with D channels, the task of TSAD is to predict the anomaly label using an anomaly detector $F(\cdot)$. During inference, the anomaly detector computes an anomaly score a_t for each time step t based on a sliding window of inputs, formulated as: $a_t = F(\mathbf{x}_{t-T:t})$, where T denotes the window size. The predicted anomaly label $\hat{y}_t$ is then obtained by thresholding the anomaly score: $\hat{y}_t = 1$ if $a_t > a_{\text{th}}$, and $\hat{y}_t = 0$ otherwise, where a_{th} is a predefined threshold.

3.2 The Mamba Model

The core model in Mamba is the S6, which extends conventional SSM by incorporating time-dependent parameters. It can be formulated as follows:

$$y_t = \text{SSM}(x_t, \mathbf{A}, \mathbf{B}_t, \mathbf{C}_t, \Delta_t), \tag{1}$$

where the SSM process $\text{SSM}(\cdot)$ is defined by:

$$\mathbf{h}_t = \overline{\mathbf{A}}_t \odot \mathbf{h}_{t-1} + \overline{\mathbf{B}}_t \cdot x_t, \tag{2}$$

$$y_t = \mathbf{C}_t^T \cdot \mathbf{h}_t \tag{3}$$

with $\odot$ denoting the Hadamard product, $x_t \in \mathbb{R}$ representing the input at t-th time step, $\mathbf{h}_t \in \mathbb{R}^N$ the hidden state, and $y_t \in \mathbb{R}$ the output. The key distinction from conventional SSMs lies in the time-dependent and input-adaptive parameters: $\overline{\mathbf{A}}_t \in \mathbb{R}^N$, $\overline{\mathbf{B}}_t \in \mathbb{R}^N$, and $\mathbf{C}_t \in \mathbb{R}^N$ which are dynamically computed at each time step based on the current input x_t and a learnable step-size $\Delta_t \in \mathbb{R}$:

$$\Delta_t = \text{Softplus}(\mathbf{W}_t^D \cdot x_t) \tag{4}$$

$$\overline{\mathbf{A}}_t = \exp(\Delta_t \cdot \mathbf{A}), \tag{5}$$

$$\overline{\mathbf{B}}_t = \Delta_t \cdot \mathbf{W}_t^B \cdot x_t, \tag{6}$$

$$\mathbf{C}_t = \mathbf{W}_t^C \cdot x_t, \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^N$, $\mathbf{W}^{\mathbf{B}} \in \mathbb{R}^N$, $\mathbf{W}^{\mathbf{C}} \in \mathbb{R}^N$, and $\mathbf{W}^{\mathbf{D}} \in \mathbb{R}$ are trainable parameters. $\text{Softplus}(\cdot)$ denotes the Softplus activation function.

The time-dependent nature of these parameters, particularly the step-size Δ_t, enables S6 to dynamically regulate the influence of current versus historical information. Specifically, when $\Delta_t \to 0$, the model downweights the current input, emphasizing historical context. Conversely, a larger Δ_t increases the influence of the current input, allowing the model to focus on new information. It is worth noting that the above formulation describes the univariate case. For multivariate time series, this mechanism can be naturally extended by generating Δ_t, $\overline{\mathbf{A}}_t$, and $\overline{\mathbf{B}}_t$ in a channel-independent manner, while sharing $\mathbf{C}_t$ across all channels.

3.3 Hidden Attention Map in Mamba

As noted in [Ali *et al.*, 2025], Mamba can provide a hidden attention map analogous to the SA mechanism, thereby enhancing the interpretability of its internal representations. Assuming the initial hidden state $\mathbf{h}_0 = \mathbf{0}$, Eq. (2)-(3) can be unrolled as follows:

$$\mathbf{h}_t = \sum_{i=1}^{t} \Big(\prod_{j=i+1}^{t} \overline{\mathbf{A}}_j \Big) \odot \overline{\mathbf{B}}_i \cdot x_i, \tag{8}$$

$$\begin{aligned} y_t &= \mathbf{C}_t^T \cdot \mathbf{h}_t \\ &= \mathbf{C}_t^T \cdot \sum_{i=1}^{t} \Big(\prod_{j=i+1}^{t} \overline{\mathbf{A}}_j \Big) \odot \overline{\mathbf{B}}_i \cdot x_i. \end{aligned} \tag{9}$$

This unrolled formulation can be expressed in matrix form as:

$$[y_1, y_2, \cdots, y_t]^T = \tilde{\boldsymbol{\alpha}}_t \cdot [x_1, x_2, \cdots, x_t]^T, \tag{10}$$

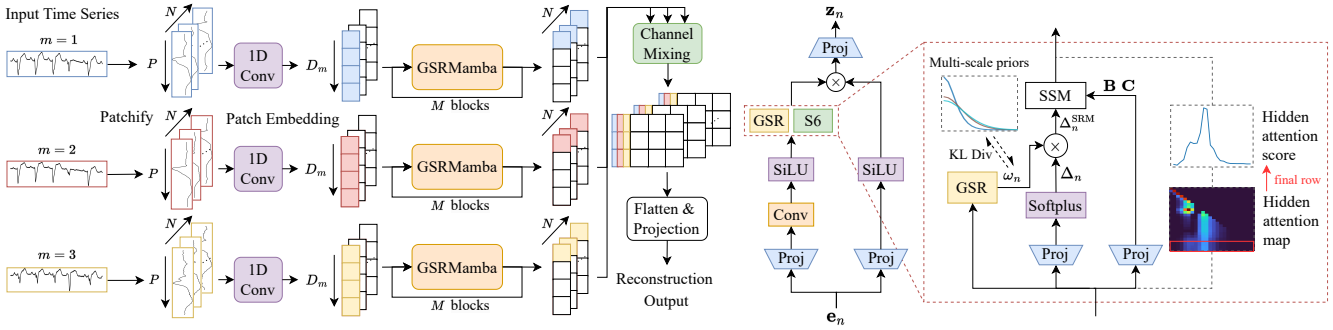
where

$$\tilde{\boldsymbol{\alpha}}_t = \begin{bmatrix} \mathbf{C}_1 \overline{\mathbf{B}}_1 & 0 & \cdots & 0 \\ \mathbf{C}_2 \overline{\mathbf{A}}_2 \odot \overline{\mathbf{B}}_1 & \mathbf{C}_2 \overline{\mathbf{B}}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_t \prod_{j=2}^{t} \overline{\mathbf{A}}_j \odot \overline{\mathbf{B}}_1 & \mathbf{C}_t \prod_{j=3}^{t} \overline{\mathbf{A}}_j \odot \overline{\mathbf{B}}_2 & \cdots & \mathbf{C}_t \overline{\mathbf{B}}_t \end{bmatrix}$$

Clearly, $\tilde{\boldsymbol{\alpha}}_t$ plays a role similar to the attention weights in Transformer models, providing interpretable attribution over the input sequence. Specifically, the contribution of input x_i to the output y_t ($t \geq i$) can be written as:

$$\mathbf{C}_t \prod_{j=i+1}^{t} \overline{\mathbf{A}}_j \odot \overline{\mathbf{B}}_i = \mathbf{C}_t \cdot \exp(\mathbf{A} \cdot \sum_{j=i+1}^{t} \Delta_j) \odot (\Delta_i \cdot \mathbf{B}_i). \tag{11}$$

This perspective endows Mamba with self-attention–like interpretability and reveals how the influence of step-size Δ_t.



(a) The process in AnoMamba. (b) GSRMamba. (c) GSR guided S6.

Figure 2: The overall framework of the proposed AnoMamba.

4 Methodology

To align the reconstruction with anomaly detection, a novel detector named AnoMamba is proposed. The overall framework is illustrated in Fig. 2, and its core design principles can be summarized in two key aspects: First, AnoMamba utilizing channel independent patch embedding to disentangle the channel dependencies and reduce the local dependency redundancy. Second, the GSRMamba blocks fully unleash the selective property of Mamba to capture appropriate global dependencies, facilitating normal behavior learning. Moreover, it offers an interpretability interface via the hidden attention map, which reveals what normal behavior has been learned by the model. In the following sections, we provide a detailed description of the AnoMamba architecture and its training objective.

4.1 Patch Embedding Layer

Given a D-channel input time series $\mathbf{x}_t \in \mathbb{R}^D$, a channel-independent patch embedding layer is first applied to capture local dependencies. Specifically, the input $\mathbf{x}_t$ is segmented into N patches, where the n-the patch $\mathbf{s}_n \in \mathbb{R}^P$ is defined as:

$$\mathbf{s}_n = \mathbf{x}_{(t+nS):(t+nS+P)}, \tag{12}$$

with P and S denoting the patch size and patch stride, respectively. Each patch is then processed by a 1D convolutional layer $\mathrm{Conv}(\cdot)$ to obtain the corresponding patch embedding $\{\mathbf{e}_n\}_{n=1}^N \in \mathbb{R}^{D_m}$, where $\mathbf{e}_n = \mathrm{Conv}(\mathbf{s}_n)$.

4.2 GSRMamba

The patch embedding has reduced the local dependencies redundancy by intra-patch modeling. However, when trained solely with a reconstruction loss, the model remains prone to overfitting, and global dependency tends to be gradually discarded. To address this issue, GSRMamba is proposed to continuously track global dependencies. The architecture of GSRMamba is depicted in Fig. 2 (b), and its processing steps can be expressed as follows:

$$\mathbf{u}_n = \mathrm{SiLU}(\mathrm{Conv}(\mathbf{W}_1 \cdot \mathbf{e}_n)), \tag{13}$$

$$\mathbf{g}_n = \mathrm{SiLU}(\mathbf{W}_2 \cdot \mathbf{e}_n), \tag{14}$$

$$\mathbf{y}_n^{\mathrm{SSM}} = \mathrm{SSM}(\mathbf{u}_n, \mathbf{A}, \mathbf{B}_n, \mathbf{C}_n, \mathbf{\Delta}_n^{\mathrm{GSR}}) \tag{15}$$

$$\mathbf{z}_n = \mathbf{W}_3 \cdot (\mathbf{y}_n^{\mathrm{SSM}} \odot \mathbf{g}_n), \tag{16}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{D_m \times 2D_m}$, $\mathbf{W}_3$ are trainable projection matrices, $\mathrm{SiLU}(\cdot)$ denotes the SiLU activation function. The process in Eq. (15) denotes a reweighted step-size guided S6, which is shown in Fig. 2 (c). The reweighted step-size computed by: $\mathbf{\Delta}_n^{\mathrm{GSR}} = \mathbf{\Delta}_n \odot \boldsymbol{\omega}_n \in \mathbb{R}^{2D_m}$, where $\boldsymbol{\omega}_n$ is the reweighting coefficient. The step-size reweighting process is a central design in GSRMamba for global dependencies capturing, which will be detailed next.

4.3 Global Step-size Reweighting Model

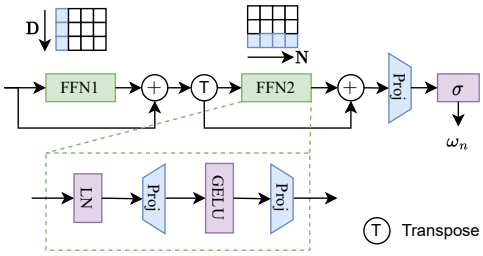


Figure 3: The architecture of the proposed GSR model.

As discussed above, the input-dependent step-size in Mamba has the potential to select reliable global dependencies. However, the step-size in original S6 depends only on the current time step, making it vulnerable to local overfitting. To exploit global information, we introduce a Global Step-size Reweighting (GSR) model that generates coefficients $\boldsymbol{\omega}_{1:N}$ from the global context and uses them to reweight the step-size. The architecture of GSR is illustrated in Fig. 3 and defined as:

$$\mathbf{F}_1 = \mathrm{FFN1}(\mathbf{u}_{1:N}) + \mathbf{u}_{1:N}, \tag{17}$$

$$\mathbf{F}_2 = \mathrm{FFN2}(\mathbf{F}_1^T)^T + \mathbf{F}^1, \tag{18}$$

$$\boldsymbol{\omega}_{1:N} = \mathrm{Sigmoid}(\mathbf{W}_4 \cdot \mathbf{F}_2), \tag{19}$$

where $\mathrm{FFN1}(\cdot)$ and $\mathrm{FFN2}(\cdot)$ are two Feed-Forward Networks (FFN) to capture intra- and inter-patch information. The resulting coefficients $\boldsymbol{\omega}_n$ thus encode global context and provide a more reliable estimate of the importance for each time step. By reweighting the original step-sizes to $\mathbf{\Delta}_n^{\mathrm{GSR}}$, the model focuses more on globally consistent dependencies and alleviates local overfitting.

4.4 Multi-scale Priors Regularization

To further prevent trivial reconstructions under unsupervised settings, we propose multi-scale priors to guide GSR. Intuitively, we expect larger step-sizes for global dependencies and smaller but non-zero step-sizes for local ones, so that global information is emphasized without discarding local context. To fulfill this, we design a family of Gamma-like long-tail distributions:

$$P(n) = (n+\alpha)^{\alpha} \cdot \exp(-n), \tag{20}$$

where n indexes the time step and $\alpha \geq 1$ controls the trade-off between global and local importance. A smaller α places more emphasis on global dependencies. The derivation of these distributions is provided in the appendix.

Motivated by multi-scale modeling, we adopt multi-scale priors by using multiple values of α instead of a single scale. To avoid hyperparameter tuning and reduce computational overhead, we randomly sample $\alpha \in [1, T]$ at each training iteration to guide the coefficients generated by GSR. As formulated in Eq. (11), the prior on step-sizes only biases how strongly global information is propagated, while the hidden attention scores are still determined jointly by the learnable parameters $\mathbf{A}$, $\mathbf{B}_t$, and $\mathbf{C}_t$. Thus, unlike attention priors, it does not impose any positional assumptions.

4.5 Reconstruction Output

After being processed by multiple GSRMamba blocks, the channel dependencies are finally mixed using an FFN model:

$$\mathbf{z}^{C}_{n,d} = \mathrm{FFN}_3(\{\mathbf{z}_{n,d}\}^{D}_{d=1}). \tag{21}$$

After flattening and projection, the final reconstructed output $\hat{\mathbf{x}}_t$ is obtained as:

$$\{\hat{\mathbf{x}}_{t,d}\}^{T}_{t=1} = \mathbf{W}^{\mathbf{O}} \cdot \mathrm{Flatten}(\{\mathbf{z}^{C}_{n,d}\}^{N}_{n=1}), \tag{22}$$

where $\mathbf{W}^{\mathbf{O}} \in \mathbb{R}^{T \times N \cdot D_m}$ is a trainable parameter matrix.

4.6 Training Objective and Anomaly Score

During training, the objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda \cdot \mathcal{L}_{\mathrm{KL}}, \tag{23}$$

where $\mathcal{L}_{\mathrm{rec}} = \mathrm{MSE}(\mathbf{x}_t, \hat{\mathbf{x}}_t)$ is the Mean Squared Error (MSE) between the original input and the reconstructed output, and $\mathcal{L}_{\mathrm{KL}}$ is the averaged Kullback–Leibler (KL) divergence between the reweighting coefficients and the prior. In each block and for the i-th dimension, the KL term is computed as $\sum_{n=1}^{T} \omega_{i,n}\big(\log \omega_{i,n} - \log P(n)\big)$. The hyperparameter λ balances the two loss terms.

During inference, the reconstruction error is used as the anomaly scores: $a_t = \mathcal{L}_{\mathrm{rec}}(\mathbf{x}_t, \hat{\mathbf{x}}_t)$.

5 Experiments

In this section, we provide the experimental setups, results, and the corresponding conclusions.

5.1 Datasets

Experiments were conducted on eight real-world benchmark datasets, including four univariate datasets from UCR Anomaly [Wu and Keogh, 2021] (ECG, EPG, Gait, NASA) and four multivariate datasets from TSB-AD [Liu and Paparrizos, 2024] (LTDB, MITDB, SVDB, SMD).

5.2 Baselines

Our method was compared against representative and SOTA TSAD methods, including: (i) MLP-based methods: Dlinear [Zeng *et al.*, 2023], TimeMixer [Wang *et al.*, 2024b], and LTFAD [Chen *et al.*, 2025]. (ii) CNN-based methods: TimesNet [Wu *et al.*, 2023], MICN [Wang *et al.*, 2023], ModernTCN [Luo and Wang, 2024]. (iii) RNN-based methods: Omnianomaly [Su *et al.*, 2019]. (iv) Transformer-based methods: Autoformer [Wu *et al.*, 2021], AnomalyTrans [Xu *et al.*, 2021], PatchTST [Nie *et al.*, 2023], D3R [Wang *et al.*, 2024a], and RTdetector [Liu *et al.*, 2025] (v) Mamba-based methods: SST [Xu *et al.*, 2024], MS-Mamba [Karadag *et al.*, 2025] and DMamba [Chen *et al.*, 2024].

5.3 Implemented Details

For fair comparison, the window size of all baselines was set to 100, following recommendations from most prior works. As noted by [Kim *et al.*, 2022], point-adjustment metrics can lead to misleading rankings, we adopt two alternative evaluation metrics: affiliation-based F1 score (F1-AF) [Huet *et al.*, 2022] and the volume under the ROC surface (VUS-ROC) [Paparrizos *et al.*, 2022]. Higher values in both metrics indicate better anomaly detection performance. All thresholds were selected by SPOT [Siffer *et al.*, 2017]. For AnoMamba, we set the model size to 16, the state size to 8, and the number of blocks to 3. All experiments were conducted in PyTorch on an NVIDIA P40 24GB GPU, with results averaged over five seeds.

5.4 Main Results

Tab. 1 and Tab. 2 report the results on univariate and multivariate datasets. The best results are **bolded** and the second-best are underlined. AnoMamba achieves the best F1-AF on $7/8$ datasets, the best VUS-ROC on $6/8$ datasets, and the best overall averaged performance. Notably, TimeMixer performs well on Gait and NASA due to its multi-resolution views, but degrades on other datasets. Similarly, MS-Mamba works better on multivariate datasets but struggles on univariate ones, suggesting high sensitivity to hyperparameters and limited generalizability. Transformer-based methods perform relatively poorly, especially AnomalyTrans, whose attention priors impose overly strong positional assumptions. Other Mamba-based methods, such as SST, are also unstable, especially on Gait with many contextual anomalies. These results show that existing Mamba-based methods still suffer from overfitting in unsupervised TSAD, while AnoMamba successfully resolves this issue and consistently achieves robust performance across diverse datasets and settings.

5.5 Ablation Studies

Tab. 3 evaluates each component of AnoMamba and several conclusions can be drawn. First, removing the GSR model leads to a significant performance drop, especially on univariate datasets dominated by contextual anomalies, underscoring the importance of reweighting in AnoMamba. Second, with fixed priors, the performance is insensitive to the choice of α, showing that the designed prior guides step-sizes to capture

Method	ECG		EPG		Gait		NASA		Avg.	
	F1-AF	VUS-ROC	F1-AF	VUS-ROC	F1-AF	VUS-ROC	F1-AF	VUS-ROC	F1-AF	VUS-ROC
Dlinear [Zeng *et al.*, 2023]	0.5623	0.6255	0.6521	0.6333	0.6249	0.6963	0.7038	0.8014	0.6460	0.6805
TimeMixer [Wang *et al.*, 2024b]	0.5476	0.7300	0.6674	0.6263	0.8245	**0.8055**	0.7095	**0.8166**	0.6939	0.7162
LTFAD [Chen *et al.*, 2025]	0.3407	0.6551	0.3031	0.5237	0.5289	0.6342	0.2062	0.6188	0.3347	0.5830
TimesNet [Wu *et al.*, 2023]	0.6481	0.7335	0.6504	0.5868	0.4910	0.6303	0.6598	0.7354	0.6189	0.6452
MICN [Wang *et al.*, 2023]	0.7364	0.6778	0.7032	0.6547	0.5802	0.6891	0.7563	0.7335	0.6928	0.6812
ModernTCN [Luo and Wang, 2024]	0.7510	0.7731	0.6983	0.5979	0.7265	0.6769	0.6898	0.6790	0.7090	0.6532
Omnianomaly [Su *et al.*, 2019]	0.6831	0.5986	0.6662	0.4572	0.4006	0.4848	0.6124	0.5428	0.6018	0.4985
Autoformer [Wu *et al.*, 2021]	0.6808	0.6486	0.6241	0.6235	0.6622	0.5696	0.7330	0.6690	0.6618	0.6249
AnomalyTrans [Xu *et al.*, 2021]	0.6654	0.5015	0.6648	0.5008	0.6678	0.5023	0.6667	0.5029	0.6659	0.5016
PatchTST [Nie *et al.*, 2023]	0.7180	0.7410	0.5946	0.7008	0.6885	0.7628	0.6807	0.7957	0.6475	0.7385
D3R [Wang *et al.*, 2024a]	0.6504	0.6513	0.6647	0.5583	0.6693	0.6907	0.6603	0.6524	0.6630	0.6171
RTdetector [Liu *et al.*, 2025]	0.5915	0.6971	0.7155	0.7014	0.5275	0.5724	0.5289	0.6446	0.6220	0.6621
SST [Xu *et al.*, 2024]	0.7441	0.8064	0.6861	0.6180	0.7558	0.6790	0.7632	0.7935	0.7239	0.6908
MS-Mamba [Karadag *et al.*, 2025]	0.6010	0.6892	0.7045	0.6738	0.6085	0.6345	0.7534	0.7511	0.6818	0.6836
DMamba [Chen *et al.*, 2024]	0.6471	0.6414	0.7074	0.7328	0.7414	0.7393	0.7054	0.7682	0.7065	0.7301
AnoMamba (Proposed)	**0.7984**	**0.8209**	**0.7514**	**0.7374**	**0.8312**	0.8046	**0.7834**	0.7953	**0.7830**	**0.7739**

Table 1: Experimental results on Univariate Datasets

Method	LTDB		MITDB		SVDB		SMD		Avg.	
	F1-AF	VUS-ROC	F1-AF	VUS-ROC	F1-AF	VUS-ROC	F1-AF	VUS-ROC	F1-AF	VUS-ROC
Dlinear [Zeng *et al.*, 2023]	0.6624	0.6553	0.7046	0.5851	0.6866	0.6577	0.6998	0.8423	0.6929	0.7124
TimeMixer [Wang *et al.*, 2024b]	0.7012	0.6706	0.7302	0.6552	0.7385	0.6969	0.7294	0.8798	0.7314	0.7547
LTFAD [Chen *et al.*, 2025]	0.6865	0.6547	0.6751	0.5338	0.6694	0.6270	0.6756	0.5110	0.6737	0.5709
TimesNet [Wu *et al.*, 2023]	0.7323	0.7123	0.7211	0.6950	0.7476	0.7213	0.8175	0.8931	0.7676	0.7787
MICN [Wang *et al.*, 2023]	0.8040	0.6802	0.7420	0.6468	0.7745	0.6727	0.7319	0.8819	0.7554	0.7449
ModernTCN [Luo and Wang, 2024]	0.7565	0.7088	0.7503	0.7026	0.7509	0.7282	0.7157	0.8950	0.7384	0.7833
Omnianomaly [Su *et al.*, 2019]	0.7226	0.5902	0.5125	0.5145	0.6574	0.5485	0.3418	0.6038	0.5224	0.5656
Autoformer [Wu *et al.*, 2021]	0.6585	0.6251	0.7006	0.6161	0.7329	0.6341	0.7326	0.8823	0.7225	0.7207
AnomalyTrans [Xu *et al.*, 2021]	0.6983	0.7002	0.6322	0.6918	0.6700	0.6839	0.6678	0.5058	0.6647	0.6215
PatchTST [Nie *et al.*, 2023]	0.7017	0.7026	0.7131	0.6649	0.7323	0.7093	0.8218	0.8947	0.7596	0.7688
D3R [Wang *et al.*, 2024a]	0.7576	0.6946	0.7305	0.6720	0.7221	0.6559	0.7994	0.8562	0.7539	0.7340
RTdetector [Liu *et al.*, 2025]	0.6695	0.6834	0.7067	0.6778	0.7093	0.6347	0.8072	0.8350	0.7419	0.7180
SST [Xu *et al.*, 2024]	**0.8097**	0.7207	0.6194	0.6672	0.7525	0.7364	0.7911	0.8889	0.7478	0.7792
MS-Mamba [Karadag *et al.*, 2025]	0.7811	0.6841	0.7365	0.6750	0.7630	0.6694	0.7979	0.8346	0.7724	0.7314
DMamba [Chen *et al.*, 2024]	0.7496	0.6446	0.7139	0.5793	0.7345	0.6313	0.8015	0.8729	0.7564	0.7112
AnoMamba (Proposed)	0.7801	**0.7369**	**0.7531**	**0.7045**	**0.7762**	**0.7545**	**0.8242**	**0.8985**	**0.7900**	**0.7973**

Table 2: Experimental results on Multivariate Datasets

Method	Univariate dataset		Multivariate dataset	
	F1-AF	VUS-ROC	F1-AF	VUS-ROC
AnoMamba	**0.7830**	**0.7739**	**0.7900**	**0.7973**
w/o GSR	0.6547	0.7425	0.7764	0.7701
fix prior $\alpha = 5$	0.7810	0.7721	0.7804	0.7911
fix prior $\alpha = 9$	0.7818	0.7734	0.7778	0.7917
fix prior $\alpha = 13$	0.7804	0.7718	0.7840	0.7913

Table 3: Ablation Studies

appropriate dependencies without introducing strong assumptions. Finally, multi-scale priors further enhances generalizability.

5.6 Sensitivity Analysis

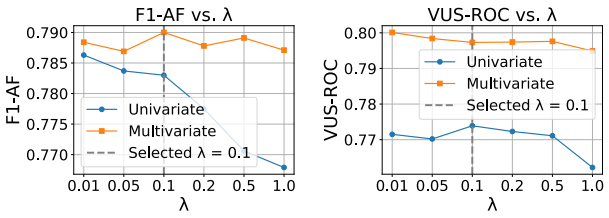
Fig. 4 shows the sensitivity of AnoMamba to key hyperparameters. For λ, overly large values cause a sharp drop in F1-AF, while performance is stable when $\lambda < 0.1$. Based on this observation, we set $\lambda = 0.1$ in our experiments. As for the patch size, performance remains within a reasonable range, and we choose a smaller size of 8 to prioritize F1-AF and better exploit GSRMamba's global information selection.

5.7 Efficiency Analysis

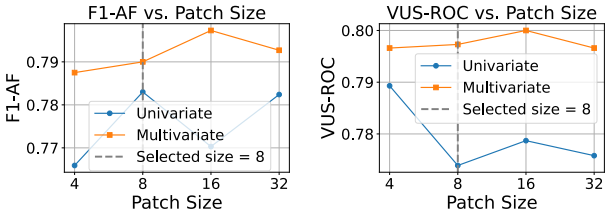
To evaluate the efficiency of AnoMamba, we analyze training time and memory usage on the MITDB dataset, comparing it with representative baselines, as presented in Fig. 5. AnoMamba demonstrates higher efficiency than multi-scale based methods, such as TimeMixer and MS-Mamba, and Transformer-based methods, such as Autoformer. Its efficiency is comparable to most CNN-based approaches, while achieving superior detection performance. Although AnoMamba is less efficient than DLinear, it significantly outperforms DLinear in detection performance.

5.8 Interpretability Analysis

To better understand the role of GSR in AnoMamba, Fig. 6 visualizes the hidden attention scores on the "GP711MarkerLFM5z1" dataset, with and without GSR guidance. It shows that vanilla Mamba overfits to local patterns, while applying GSR effectively captures global dependencies and alleviates local overfitting. Fig. 7 further visualizes the corresponding anomaly scores. The input sequence in the first row exhibits a global pattern characterized by a "large peak–small peak" structure, where the anomaly is caused by



(a) Results with different λ values.



(b) Results with different patch sizes.

Figure 4: Sensitivity analysis.

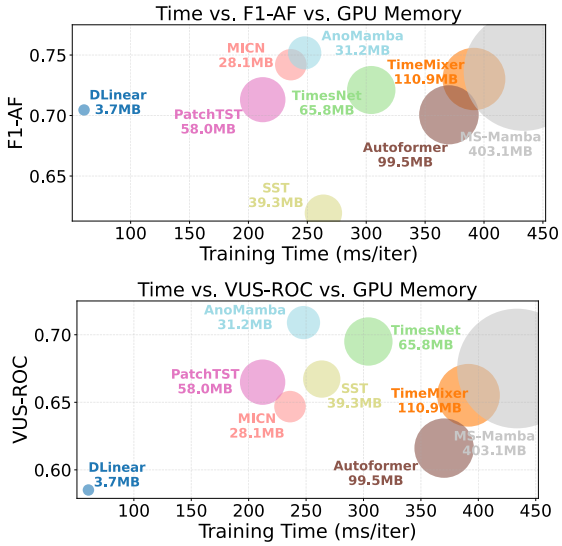


Figure 5: Visualization of the computational complexity.

the absence of the "small peak". In the second and third rows, we observe that under GSR guidance, the reconstruction of the anomalous region attends to the "large peak," indicating that the model has successfully learned the global dependencies of the normal behavior. This allows AnoMamba to precisely detect the anomaly via reconstruction error. In contrast, the fourth and fifth rows show that without GSR, Mamba tends to overfit local dependencies, resulting in false negatives. These visualizations demonstrate that AnoMamba can effectively learn global normal behavior to realign reconstruction with anomaly detection.

6 Conclusions

In this paper, we identify a critical issue in reconstruction-based TSAD, namely the objective misalignment caused by

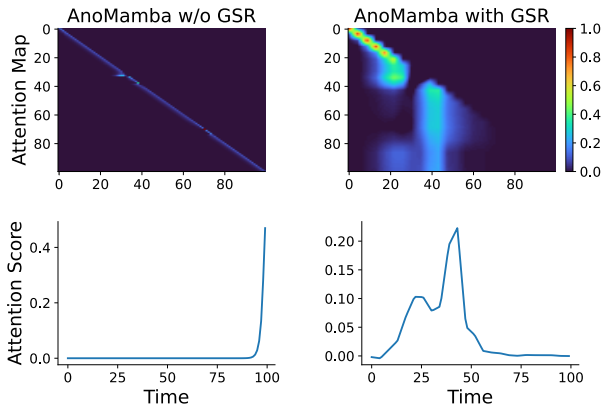


Figure 6: Visualization of hidden attention maps in AnoMamba. The last row in each map shows the hidden attention scores used for the current reconstruction step.

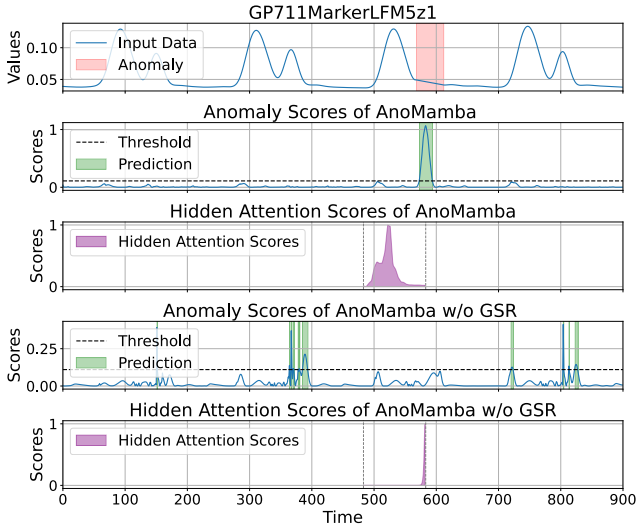


Figure 7: Visualization of anomaly scores produced by AnoMamba.

local overfitting in unsupervised settings. To address this issue, we propose AnoMamba, a novel detector that aligns the reconstruction and anomaly detection objectives. The key design centers on GSRMamba with a step-size reweighting process, which enhances global dependency modeling. To further alleviate local overfitting in the unsupervised settings, a multi-scale prior strategy is designed to guide the reweighting coefficients. Extensive experiments demonstrate both the effectiveness and efficiency of AnoMamba across diverse datasets. In addition, AnoMamba provides interpretability by revealing learned normal behavior through its hidden attention map.

Acknowledgments

The work was supported by Shenzhen Science and Technology Program under Grants KCXFZ20240903094406009 and JCYJ20230807145600001, and NSFC Grant No. 62192713.

References

[Ali *et al.*, 2025] Ameen Ali Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of mamba models. In *Proc. 63rd Annu. Meet. Assoc. Comput. Linguist.*, pages 1516–1534, 2025.

[An and Cho, 2015] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Spec. Lect. on IE*, 2(1):1–18, 2015.

[Boniol *et al.*, 2024] Paul Boniol, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Paparrizos. Dive into time-series anomaly detection: A decade review. *arXiv preprint arXiv:2412.20512*, 2024.

[Chen *et al.*, 2024] Junqi Chen, Xu Tan, Sylwan Rahardja, Jiawei Yang, and Susanto Rahardja. Joint selective state space model and detrending for robust time series anomaly detection. *IEEE Signal Process. Lett.*, 31:2050–2054, 2024.

[Chen *et al.*, 2025] Lei Chen, Xinzhe Cao, Tingqin He, Yepeng Xu, Xuxin Liu, et al. A lightweight all-mlp time–frequency anomaly detection for iiot time series. *Neural Networks*, 187:107400, 2025.

[Gu and Dao, 2024] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proc. Conf. Lang. Model.*, 2024.

[Hoffmann, 2007] Heiko Hoffmann. Kernel pca for novelty detection. *Pattern Recognit.*, 40(3):863–874, 2007.

[Huet *et al.*, 2022] Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, pages 635–645, 2022.

[Karadag *et al.*, 2025] Yusuf Meric Karadag, Sinan Kalkan, and Ipek Gursel Dino. ms-mamba: Multi-scale mamba for time-series forecasting. *arXiv preprint arXiv:2504.07654*, 2025.

[Kim *et al.*, 2022] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proc. AAAI Conf. Artif. Intell.*, volume 36, pages 7194–7201, 2022.

[Liu and Paparrizos, 2024] Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Proc. Adv. Neural Inf. Process. Syst.*, 37:108231–108261, 2024.

[Liu *et al.*, 2025] Xinhong Liu, Xiaoliang Li, Yangfan Li, Fengxiao Tang, and Ming Zhao. Rtdetector: Deep transformer networks for time series anomaly detection based on reconstruction trend. In *Proc. Int. Joint Conf. Artif. Intell.*, 2025.

[Luo and Wang, 2024] donghao Luo and xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *Proc. Int. Conf. Learn. Represent.*, 2024.

[Nie *et al.*, 2023] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Proc. Int. Conf. Learn. Represent.*, 2023.

[Paparrizos *et al.*, 2022] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron Elmore, and Michael J Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. *Proc. VLDB Endow.*, 15(11):2774–2787, 2022.

[Peng *et al.*, 2025] Furong Peng, Rongxin Ma, Xuan Lu, Yuhua Qian, Yong Xu, Zhiguo Hu, and Hongtao Wu. Mgrd: Multi-granularity reconstruction deviation modeling for time series anomaly detection. *IEEE Trans. Instrum. Meas.*, 2025.

[Schmidl *et al.*, 2022] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: a comprehensive evaluation. *Proc. VLDB Endow.*, 15(9):1779–1797, 2022.

[Siffer *et al.*, 2017] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. In *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, pages 1067–1075, 2017.

[Su *et al.*, 2019] 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 *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, pages 2828–2837, 2019.

[Wang *et al.*, 2023] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *Proc. Int. Conf. Learn. Represent.*, 2023.

[Wang *et al.*, 2024a] Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pages 10758–10774, 2024.

[Wang *et al.*, 2024b] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *Proc. Int. Conf. Learn. Represent.*, 2024.

[Wu and Keogh, 2021] Renjie Wu and Eamonn J Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Trans. Knowl. Data Eng.*, 35(3):2421–2429, 2021.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pages 22419–22430, 2021.

[Wu *et al.*, 2023] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *Proc. Int. Conf. Learn. Represent.*, 2023.

[Xu *et al.*, 2021] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *Proc. Int. Conf. Learn. Represent.*, 2021.

[Xu *et al.*, 2024] Xiongxiao Xu, Canyu Chen, Yueqing Liang, Baixiang Huang, Guangji Bai, Liang Zhao, and Kai Shu. Sst: Multi-scale hybrid mamba-transformer experts for long-short range time series forecasting. *arXiv preprint arXiv:2404.14757*, 2024.

[Zamanzadeh Darban *et al.*, 2024] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Comput. Surv.*, 57(1):1–42, 2024.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pages 11121–11128, 2023.

[Zhong *et al.*, 2025] Guojin Zhong, Jin Yuan, Zhiyong Li, Long Chen, et al. Multi-resolution decomposable diffusion model for non-stationary time series anomaly detection. In *Proc. Int. Conf. Learn. Represent.*, 2025.