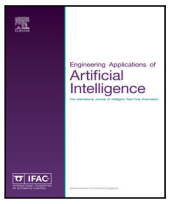
Engineering Applications of Artificial Intelligence 160 (2025) 111685

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Mamba Adaptive Anomaly Transformer with association discrepancy for time series

Abdellah Zakaria Sellam [a,d], Ilyes Benaissa [b,a], Abdelmalik Taleb-Ahmed [c], Luigi Patrono [a,d], Cosimo Distante [a,d,*]

[a] Department of Innovation Engineering, University of Salento, via Monteroni, Lecce, 73100, Italy
[b] VSC Laboratory, Department of Electrical Engineering, University of Mohamed Khider Biskra, Biskra, 07000, Algeria
[c] Institute d'Electronique de Microelectronique et de Nanotechnologie (IEMN), UMR 8520, Universite Polytechnique Hauts de France, Universite de Lille, CNRS, Valenciennes, 59313, France
[d] Institute of Applied Sciences and Intelligent Systems - CNR, via Monteroni, Lecce, 73100, Italy

ARTICLE INFO

Keywords:
Transformer
Association discrepancy
Gated attention
Mamba state space model
Sparse attention
Anomaly detection
Unsupervised learning

ABSTRACT

Anomaly detection in time series poses a critical challenge in industrial monitoring, environmental sensing, and infrastructure reliability, where accurately distinguishing anomalies from complex temporal patterns remains an open problem. While existing methods, such as the Anomaly Transformer leveraging multi-layer association discrepancy between prior and series distributions and Dual Attention Contrastive Representation Learning architecture (DCdetector) employing dual-attention contrastive learning, have advanced the field, critical limitations persist. These include sensitivity to short-term context windows, computational inefficiency, and degraded performance under noisy and non-stationary real-world conditions. To address these challenges, we present MAAT (Mamba Adaptive Anomaly Transformer), an enhanced architecture that refines association discrepancy modeling and reconstruction quality for more robust anomaly detection. Our work introduces two key contributions to the existing Anomaly transformer architecture: Sparse Attention, which computes association discrepancy more efficiently by selectively focusing on the most relevant time steps. This reduces computational redundancy while effectively capturing long-range dependencies critical for discerning subtle anomalies. A Mamba-Selective State Space Model (Mamba-SSM) is also integrated into the reconstruction module. A skip connection bridges the original reconstruction and the Mamba-SSM output, while a Gated Attention mechanism adaptively fuses features from both pathways. This design balances fidelity and contextual enhancement dynamically, improving anomaly localization and overall detection performance. Extensive experiments on benchmark datasets demonstrate that MAAT significantly outperforms prior methods, achieving superior anomaly distinguishability and generalization across diverse time series applications. By addressing the limitations of existing approaches, MAAT sets a new standard for unsupervised time series anomaly detection in real-world scenarios. Code available at https://github.com/ilyesbenaissa/MAAT.

1. Introduction

Anomaly detection in time series data is critical in finance, healthcare, industrial monitoring, and cybersecurity to prevent failures, detect fraud, and ensure efficiency. Traditionally, methods like ARIMA (Box and Jenkins, 1970) and Gaussian Processes (Rasmussen and Williams, 2006) have been used to identify abnormal events in sequential data. Which relies on the assumption that anomalies can be identified as deviations from predicted values. However, these methods often struggle with real-world time series data of a complex, high-dimensional nature, especially when the underlying patterns are non-linear or when the anomalies are subtle (Box et al., 2015). Machine learning introduced advanced models like Support Vector Machines (SVMs) (Scholkopf et al., 2000), Random Forests (Rabiner and Juang, 1986), and Hidden Markov Models (HMMs) (Hochreiter and Schmidhuber, 1997) were among the first machine learning methods applied for time series anomaly detection, improving on traditional methods by learning from data. However, they still required extensive feature engineering and often missed complex temporal dependencies. Deep learning models such as Recurrent Neural Networks (RNNs) (Radford et al., 2018), and Long Short-Term Memory (LSTM) (Malhotra et al., 2016) revolutionized time series anomaly

* Corresponding author at: Institute of Applied Sciences and Intelligent Systems - CNR, via Monteroni, Lecce, 73100, Italy.
E-mail address: cosimo.distante@cnr.it (C. Distante).

https://doi.org/10.1016/j.engappai.2025.111685
Received 21 February 2025; Received in revised form 25 May 2025; Accepted 1 July 2025
Available online 25 July 2025

detection by capturing long-term dependencies and complex temporal patterns, ideal for identifying subtle deviations (Hochreiter and Schmidhuber, 1997). Additionally, the use of autoencoders and their variants, such as Variational Autoencoders (VAEs) (Xu et al., 2018) and Adversarial Autoencoders (Somepalli et al., 2021), has further advanced the field by enabling unsupervised anomaly detection through the reconstruction of standard data patterns (Kingma and Welling, 2013). Reconstruction-based methods, like autoencoders (Malhotra et al., 2016), model standard data patterns and detect anomalies through reconstruction errors, leveraging their ability to learn compressed representations (Hinton and Salakhutdinov, 2006). Similarly, recurrent neural networks (RNNs) and long short-term memory (LSTM) networks have been employed to capture temporal dependencies in sequential data during reconstruction tasks (Malhotra et al., 2016). However, reliance on reconstruction loss can cause false positives for anomalies resembling standard data. Transformers (Vaswani et al., 2017), with self-attention mechanisms, address this by modeling complex dependencies in time series (Benali et al., 2024). Enhancements like gated attention (Zhang et al., 2023), for instance, integrate gating structures to dynamically control the flow of information, enabling the model to focus on relevant temporal features and improve performance in time series forecasting tasks. Sparse attention (Guo et al., 2024) mechanisms have also been introduced to improve the efficiency of Transformers when handling long sequences. By limiting the attention computation to a subset of the input sequence, sparse Attention reduces computational complexity while maintaining the ability to capture essential dependencies. This approach is particularly beneficial for long-sequence time-series forecasting, where traditional attention mechanisms may become computationally prohibitive. Recent advancements include the Anomaly Transformer (Xu et al., 2022), which uses association discrepancy learning to differentiate normal from abnormal points without relying solely on reconstruction loss. It employs Gaussian kernels and attention mechanisms to model prior- and series-association discrepancies, improving accuracy while addressing scalability in high-dimensional datasets. Despite achieving state-of-the-art performance through a minimax strategy, it faces limitations: self-attention struggles with long-range dependencies in small windows, and noise or non-stationary patterns can increase false positives during normal fluctuations. While more efficient than traditional methods, these challenges remain. Furthermore, DCdetector (Yang et al., 2023b) simplifies anomaly detection with a dual-attention contrastive structure, eliminating complex components like Gaussian kernels or reconstruction losses. It uses contrastive learning to separate anomalies from expected points, capturing global and local dependencies through a dual-channel architecture. Its channel-independent patching reduces parameter complexity and overfitting risks, while a contrastive loss function based on Kullback–Leibler divergence enhances representation consistency. However, DCdetector faces challenges, including sensitivity to contrastive sample quality, high computational overhead from pairwise comparisons, and dual-branch parallelism. Building upon these foundations, we propose MAAT (Mamba Adaptive Anomaly Transformer with association discrepancy for time series) for anomaly detection in time series. Our approach introduces a novel block architecture that incorporates skip connections and gating mechanisms in association with the mamba block in the skip connection to improve the reconstruction capability of the anomaly transformer architecture. By combining these architectural enhancements with sparse attention mechanisms and anomaly Transformer association modeling principles, MAAT performs better in detecting anomalies across diverse datasets. The Key contributions of this work are summarized as follows:

- integrated sparse attention mechanisms to replace the standard attention mechanism in the original Anomaly Transformer, allowing scalable and efficient processing of long time series data. By enabling dynamic control over block size, MAAT achieves an optimal balance between computational efficiency and precision in anomaly detection.
- Introduced the MAMBA block to the Anomaly Transformer architecture to optimize the reconstruction of signals.
- Incorporated the MAMBA block with skip connections and gating Attention to enhance the reconstruction output and reduce the loss of the Anomaly Transformer architecture.

The subsequent sections of this document are structured as follows: Section 2 presents a comprehensive overview of relevant works, including literature and methodologies, related to time series anomaly detection. Section 3 delves into our proposed MAAT architecture, detailing its unique features and potential benefits. In Section 4, we conduct Experiments, perform data analysis, and provide a thorough evaluation of our model. Lastly, in Section 5, we draw insightful conclusions based on the following: In our experimentation, we compare our approach with prior methodologies and articulate the implications of our findings. This structure ensures a coherent and comprehensive understanding of our innovative methodology for unsupervised anomaly detection in time series data.

2. Related works

Time series anomaly detection has evolved from traditional statistical models like ARIMA and Gaussian Processes (Box et al., 2015; Rasmussen and Williams, 2006). Conventional methods like ARIMA and GP struggled with non-linear and high-dimensional data. Machine learning introduced more flexible approaches, such as (SVMs), particularly One-Class SVM (OC-SVM), which excel in novelty detection by defining a boundary around standard data and identifying outliers as anomalies (Scholkopf et al., 2000). Despite this improvement, SVMs could not capture temporal dependencies inherent in sequential data. Random Forests and Isolation Forests handled high-dimensional data better than SVMs, while Isolation Forests identified anomalies by isolating points through recursive partitioning (Liu et al., 2008). However, they focused on anomalies and were not optimized for capturing time dependencies in sequential data. RNNs and LSTMs became popular due to their ability to capture long-term dependencies in time series data. These models excelled in detecting anomalies in healthcare (e.g., ECG data) and finance (Distante et al., 2022). However, RNNs and LSTMs face limitations in scalability and can struggle with vanishing gradient problems when processing long sequences. Variational Autoencoders (VAEs) became central to unsupervised anomaly detection by reconstructing standard data patterns and flagging high-reconstruction errors as anomalies (Kingma and Welling, 2013). Adversarial Autoencoders (AAEs) and Generative Adversarial Networks (GANs) enhanced robustness through adversarial training. GANs use a generator–discriminator framework to detect anomalies based on realistic data generation. However, both GANs and AEs require large datasets and can face instability, limiting their application in some domains (Schlegl et al., 2019; Li et al., 2019). Transformer-based models have recently demonstrated significant potential in time series anomaly detection by leveraging self-attention mechanisms to capture long-range dependencies (Vaswani et al., 2017). Initially developed for NLP, these models eliminate the need for recurrent structures like RNNs or LSTMs, effectively modeling local and global patterns in time series data to detect anomalies. Transformer-based models have revolutionized time series anomaly detection, with innovations like the Anomaly Transformer introducing Association Discrepancy to compare expected and observed associations in data. Using a minimax strategy enhances the distinguishability of anomalies, outperforming prior methods across datasets (Xu et al., 2022). Similarly, AnomalyBERT employs self-supervised learning and data degradation to simulate anomalies, improving generalization without labeled data (Jeong et al., 2023). The Denoising Diffusion Mask Transformer (DDMT) integrates denoising diffusion with masking mechanisms, excelling in noisy multivariate settings (Yang et al., 2023a). Multivariate anomaly detection has also advanced with models like Informer (Zhou et al., 2021) and multi-task Transformers, which

leverage attention mechanisms to model interdependencies among variables (Zhang et al., 2021). Hybrid models combining statistical methods with deep learning enhance interpretability while maintaining flexibility. These developments address the limitations of classical met techniques as ARIMA, which struggled with high-dimensional, non-stationary data. As the field progressed, deep learning frameworks began to emerge. In 2021, the **Anomaly Transformer** (Xu et al., 2022) introduced a novel anomaly-attention mechanism that quantifies the association discrepancy between normal and abnormal points. While this approach effectively leverages self-attention to capture both local and global temporal dependencies in an unsupervised manner, its quadratic complexity poses challenges for scalability, and its underlying assumptions may not hold across all types of anomalies. Building on these advances, the **DC Detector** (Yang et al., 2023b) employs a dual attention contrastive representation learning framework. Integrating local and global Attention with a contrastive loss overcomes some of the pitfalls of reconstruction-based methods. Nonetheless, the additional architectural complexity and the need for extensive hyperparameter tuning remain nontrivial hurdles. Concurrently, selective state space models have gained traction for their efficiency. Models like **MAMBA** (Gu and Dao, 2023) use a selective scanning mechanism to model long-range dependencies linearly, making them attractive for real-time and large-scale applications. However, the trade-off is a potential flexibility reduction when modeling highly non-linear interactions. Despite these significant advances, several challenges persist. One major issue is **concept drift**, the phenomenon where the underlying distribution of data evolves. This drift renders pre-trained models ineffective and highlights the need for adaptive learning techniques that can update dynamically as new data becomes available. In addition, scalability remains a critical challenge, especially when dealing with high-frequency time series data or real-time anomaly detection in large-scale systems. Recent efforts in developing memory-efficient Transformers and distributed learning approaches (Gupta et al., 2021; Lai et al., 2021) have begun to address these scalability issues, marking promising steps toward more adaptive and robust anomaly detection systems. Overall, the evolution of unsupervised time series anomaly detection reflects a broader trajectory across complex domains. The field is steadily progressing toward more nuanced, robust, and adaptable solutions, from early statistical methods to modern deep learning architectures that harness attention mechanisms, contrastive learning, and efficient state space representations. These advancements enhance our ability to capture intricate temporal dependencies and subtle deviations without relying on labeled data and pave the way for deploying these models in real-world, dynamic environments.

3. Methodology

Consider a system that records a sequence of d-dimensional measurements at uniform time intervals. The time series data is represented by the set $\{x_1, x_2, \ldots, x_N\}$, where each $x_t \in \mathbb{R}^d$. The goal is to determine whether an observation x_t is anomalous in an unsupervised manner. Effective unsupervised time series anomaly detection relies on learning informative representations and establishing a clear discriminative criterion. The original Anomaly Transformer framework distinguishes standard patterns from anomalies by learning an association discrepancy via anomaly attention and a minimax optimization strategy. Building on this, the Mamba Adaptive Anomaly Transformer (MAAT) introduces Anomaly Sparse Attention, Mamba Blocks, and Gated Skip Connections, enhancing short-range dependency modeling and long-range temporal learning while reducing computational cost. As illustrated in Fig. 1(A), Anomaly Sparse Attention replaces dense self-attention with a two-branch mechanism. The Prior-Association Branch encodes expected dependencies using a learnable Gaussian kernel, while the Series-Association Branch applies block-wise sparse Attention to capture time series patterns adaptively. These associations guide discrepancy learning, enhancing anomaly detection performance. The Reconstruction Block, shown in Fig. 1(B), refines extracted features by stacking MAAT Blocks with LayerNorm and Feed-forward Networks (FFN). This layered design supports hierarchical feature learning, making the model more robust to diverse time series behaviors. Finally, the MAAT Block, depicted in Fig. 1(C), integrates Mamba Blocks within a Gated Attention mechanism, selectively amplifying meaningful signals while filtering out noise, and the skip connection preserves critical information while adapting the model focus. MAAT provides a scalable and expressive framework for detecting complex anomalies across various time series datasets by combining sparse Attention, state–space modeling, and gated learning mechanisms with the association discrepancy.

3.1. Background knowledge

3.1.1. Association discrepancy

The Anomaly Transformer framework employs a minimax optimization strategy that alternates between two phases to balance the learning of two complementary feature sets, P and S. This alternating approach prevents collapse in the representation of P while ensuring that S captures non-local dependencies.

Minimize phase:. As introduced in the Anomaly Transformer framework, this minimax strategy ensures that both feature sets capture complementary aspects of the input data, enhancing the overall quality of the learned representations. The hyperparameter λ is critical in balancing the network's ability to reconstruct the input with the need for diverse and informative feature learning.

3.1.2. Association-based anomaly criterion

The Anomaly Transformer framework integrates temporal pattern analysis with reconstruction fidelity through a dual-mechanism scoring system. The final anomaly score for an input $X \in \mathbb{R}^{N\times d}$ is computed as:

$$\text{AnomalyScore}(X) = \text{Softmax}\Big(-\text{AssDis}(P, S; X)\Big) \odot \|X_{i,:} - \hat{X}_{i,:}\|_2^2, \quad i = 1, \ldots, N, \tag{1}$$

Where:

- $\text{AssDis}(P, S; X)$ measures the association discrepancy between prior-association (P) and series-association (S), which is formulated as:

$$\text{AssDis}(P, S; X) = \frac{1}{L}\sum_{l=1}^{L}\left[\sum_{i=1}^{N}\text{KL}(P_{i,:}^l \| S_{i,:}^l) + \text{KL}(S_{i,:}^l \| P_{i,:}^l)\right], \tag{2}$$

  - Where KL(Kullback–Leibler) divergence, compares probability distributions of the Prior and the Series associations

- $\text{Softmax}(\cdot)$ normalizes the negative association discrepancy across time steps,
- $\odot$ represents element-wise multiplication,
- $\|X_{i,:} - \hat{X}_{i,:}\|_2^2$ quantifies the reconstruction error at each time step i.

This formulation combines temporal patterns and reconstruction accuracy for anomaly detection. While improved reconstructions reduce discrepancy, anomalies with significant deviations in reconstruction or temporal patterns still yield high scores. Probabilistic weighting highlights time steps with poor reconstruction and abnormal dependencies, enhancing robustness.

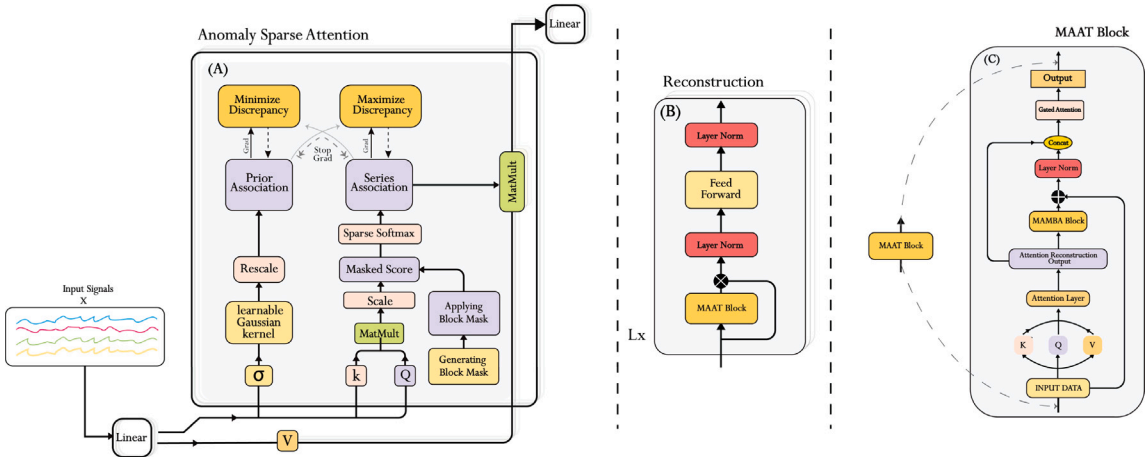


Fig. 1. The figure illustrates the MAAT framework for time series anomaly detection. Block A (Anomaly Sparse Attention Module) computes prior and series associations using sparse Attention and a learnable Gaussian kernel to model temporal dependencies. Block B (Reconstruction Module) refines the feature representations using layer normalization, feedforward processing, and MAAT blocks to reconstruct input signals effectively. Block C represents the MAAT Block that integrates the Mamba state–space model to capture long-range dependencies, followed by a gated attention mechanism that adaptively fuses the reconstructed output.

3.2. Sparse attention for efficient sequence attention

Sparse Attention restricts token interactions by applying a sparsity pattern S that selects only a subset of pairs for attention computation (Child et al., 2019; Zaheer et al., 2020; Kitaev et al., 2020). Sparse attention introduces a sparsity mask S to limit interactions:

$$\text{SparseAttention}(Q, K, V) = \text{softmax}\left(\frac{(QK^T) \odot S}{\sqrt{d_k}}\right)V, \tag{3}$$

Where:

- $Q, K, V \in \mathbb{R}^{N \times d}$ are the query, key, and value matrices,
- d_k denotes the dimensionality of the key vectors,
- The softmax function normalizes the attention scores.

where $\odot$ represents element-wise multiplication. This ensures that only selected token pairs contribute to the attention mechanism.

Approaches for defining the sparsity pattern S include:

- **Fixed Patterns:** Predefined structures, such as strided or block-wise masks (Child et al., 2019).
- **Learned Masks:** Patterns that are dynamically optimized during training (Zaheer et al., 2020).
- **Memory-Efficient Methods:** Techniques like locality-sensitive hashing (LSH) that enable content-based token selection (Kitaev et al., 2020).

The Gated Attention Mechanism enhances standard attention by using a learnable gating process to dynamically adjust the importance of input features. It evaluates the relevance of features to optimize attention weight distribution, represented mathematically as:

$$\text{GatedAttention}(x) = \sigma\big(G(x)\big) \odot A(x)$$

Key components include $x$input data, $G(x)$gating vector from a neural network, σ: sigmoid function mapping to $[0, 1]$, $A(x)$ standard attention weights, $\odot$element-wise multiplication. This mechanism allows the model to focus on task-critical features while suppressing noise, improving robustness in contexts like anomaly detection.

3.3. Mamba state space model

State Space Models (SSMs) are a class of architectures that model time series by positing a latent state evolving according to linear dynamics and observed through a readout, capturing long-term dependencies and latent structures with provable efficiency and stability (Gu et al., 2022). The **Mamba State Space Model (Mamba SSM)** builds on the Structured State Space (S4) lineage by introducing *selective* state spaces i.e., SSM parameters that vary as functions of the input—thereby enabling content-based propagation or forgetting of information along the sequence. Unlike standard subquadratic architectures, Mamba achieves *O(N)* compute and memory complexity in the sequence length N, supporting training and inference on sequences of up to $2^{20} \approx 1.05 \times 10^6$ time steps ($N = 1.05 \times 10^6$) without performance degradation. This linear scaling stems from its input-dependent parameterization and hardware-aware parallel scan algorithm, which eliminates the quadratic overhead of attention-based models. The latter also delivers up to five-fold higher throughput than same-sized Transformers in inference, making Mamba both fast and resource efficient. As a universal sequence modeling backbone, Mamba achieves state-of-the-art results across diverse modalities (language, audio, genomics)—outperforming or matching Transformers of equal or larger size in both pretraining and downstream tasks (Gu and Dao, 2023), while maintaining *O(N)* time and memory complexity for sequences of arbitrary length N.

3.4. Mamba Adaptive Anomaly Transformer Architecture

Fig. 1 illustrates the overall structure of the Mamba Adaptive Anomaly Transformer (MAAT), an architecture inspired by the Anomaly Transformer. MAAT integrates a sparse attention mechanism, a selective state–space model known as the Mamba block, and an adaptive reconstruction framework. The sparse attention mechanism processes the input time series data, capturing local dependencies efficiently. The Mamba block is designed to capture long-range dependencies, enhancing the model's ability to understand complex temporal patterns. An adaptive gating mechanism then fuses the output Attention sparse attention and Mamba block, enabling the model to reconstruct the input signal effectively. This combination allows MAAT to detect anomalies by comparing the reconstructed signal with the original input, identifying deviations that may indicate anomalous behavior.

3.4.1. Anomaly sparse attention

Our anomaly sparse attention module, shown in Fig. 1(A), refines the conventional anomaly attention used in the Anomaly Transformer, which relies on self-attention and sparse Attention. While the prior-association remains unchanged, modeled using a learnable Gaussian kernel the series-association is now computed through sparse self-attention, capturing observed dependencies more efficiently.

In this module, the prior association continues to be represented by a learnable Gaussian kernel, while the series association is computed using a sparse softmax operation over local attention windows.

For each query vector Q_i, normalization is performed only over keys in the local window defined by

$$\Omega_i = \{j \mid |j - i| \le \text{block\_size}/2\}, \tag{4}$$

so that the series association is computed as

$$S^l_{i,j} = \frac{\exp\left(\frac{Q_i K_j^T}{\sqrt{d_{\text{model}}}}\right)}{\sum_{k\in\Omega_i}\exp\left(\frac{Q_i K_k^T}{\sqrt{d_{\text{model}}}}\right)}, \quad \forall\, j \in \Omega_i, \tag{5}$$

with

$$S^l_{i,j} = 0 \quad \text{for } j \notin \Omega_i. \tag{6}$$

This sparse softmax formulation ensures that only locally relevant keys are considered during normalization, reducing computational redundancy while preserving essential dependencies.

3.4.2. MAAT adaptive block

The adaptive block shown in Fig. 1(C) integrates a state–space model (Mamba block) with a skip connection that retains the original input. The Mamba block is specifically designed to capture long-range dependencies, while the skip connection ensures the preservation of fine details. We achieve a robust intermediate representation by combining these elements and applying layer normalization. This Adaptive Block merges long-range dependencies and local features through a state–space (Mamba) skip path and an adaptive gating mechanism. After performing sparse attention processing, we proceed with the following steps: The Mamba block generates a transformed representation $x_{Mambamamba}$, which is then combined with the reconstructed input from sparse attention x_{orig} via a residual connection. This output is normalized as follows:

$$x_{\text{skip}} = \text{LayerNorm}\left(x_{\text{mamba}} + x_{\text{orig}}\right), \tag{7}$$

Adaptive gating and reconstruction:. In the context of adaptive gating and reconstruction, a gating factor g is computed from the concatenation of the output from the central processing path x and the corresponding skip connection x_{skip}. Formally, the gating factor can be expressed as follows:

$$g = \sigma\Big(\text{Linear}\big([x;\, x_{\text{skip}}]\big)\Big) = \sigma\Big(W[x;\, x_{\text{skip}}] + b\Big), \tag{8}$$

Where $[x;\, x_{\text{skip}}]$ denotes the concatenation of the feature maps along the channel dimension, W and b represent the learnable parameters of the linear transformation (precisely, the weight matrix and bias term), and $\sigma(\cdot)$ is the sigmoid activation function that maps the output to the range $(0, 1)$.

The gating factor g is critical in modulating the relative importance of the skip and main paths. Specifically, for each element:

- When g approaches 1, the output is predominantly influenced by the skip connection x_{skip}, which encapsulates long-range contextual information along with a residual signal from the original input.
- Conversely, when g approaches 0, the reconstruction is primarily derived from the main path output x, which reflects the locally processed features.

The final output, denoted as X^{adapt}, is computed as an element-wise weighted combination of these two representations:

$$X^{\text{adapt}} = g \odot x_{\text{skip}} + (1 - g) \odot x, \tag{9}$$

where $\odot$ signifies element-wise multiplication. This formulation enables the network to dynamically calibrate the contributions from the skip connection and the main processing branch, thereby facilitating a harmonious integration of local details and global context. The process of adaptive fusion is paramount in ensuring that the reconstructed output X^{adapt} accurately embodies normal patterns. In contrast, anomalies disrupt local consistency and global structure, increasing reconstruction errors. This enhanced reconstruction capability, in turn, fosters improved performance in anomaly detection.

Consequently, X^{adapt} represents a sophisticated integrated reconstruction, merging information from both the skip path and the central processing branch, ultimately enhancing the overall representational capacity of the model.

The anomaly score now leverages the adaptively fused reconstruction X^{adapt} by balancing the association discrepancy from Eq. (10) and the reconstruction error. Specifically, the anomaly score for each time step i is defined as

$$\text{AnomalyScore}(X) = \text{Softmax}\Big(-\text{AssDis}(P, S; X)\Big) \odot \Big\| X_{i,:} - X^{\text{adapt}}_{i,:} \Big\|_2^2, \quad i = 1, \ldots, N. \tag{10}$$

Here, $\odot$ represents element-wise multiplication, and:

- $\text{AnomalyScore}(X) \in \mathbb{R}^{N\times 1}$ provides the point-wise anomaly criterion.
- $\text{AssDis}(P, S; X)$ is the association discrepancy, as defined in Eq. (10).
- $\|X_{i,:} - X^{\text{adapt}}_{i,:}\|_2^2$ is the reconstruction error between the input X and its adaptively reconstructed counterpart X^{adapt}.

This updated formulation clearly distinguishes the adaptively reconstructed output X^{adapt}, which utilizes the gating for reconstruction from the input X, emphasizing the role of adaptive fusion in our anomaly detection criterion.

4. Experiments

4.1. Benchmark datasets

In this study, we evaluate the performance of our model using seven representative benchmarks derived from real-world applications. The first dataset, MSL, is the Mars Science Laboratory dataset collected by NASA, which reflects the condition of sensors and actuator data from the Mars rover (NASA, 2018a). Similarly, the SMAP dataset, provided by NASA, presents soil samples and telemetry information from the Mars rover; notably, SMAP contains more point anomalies than MSL (NASA, 2018b). The PSM dataset, a public resource from eBay Server Machines, includes 25 dimensions and is widely used for research in anomaly detection (Huang et al., 2021). Additionally, the SMD dataset consists of a five-week-long record of resource utilization traces collected from an internet company compute cluster, monitoring 28 machines (Su et al., 2019). Another critical benchmark is the SWaT dataset, which comprises 51-dimensional sensor data from a secure water treatment system that operates continuously (Mathur and Tippenhauer, 2016). Furthermore, the NIPS-TS-SWAN dataset provides a comprehensive multivariate time series benchmark extracted from solar photospheric vector magnetograms in the Spaceweather HMI Active Region Patch series (NASA, 0000). Lastly, the NIPS-TS-GECCO dataset is a drinking water quality dataset for the Internet of Things, published in the 2018 Genetic and Evolutionary Computation Conference (Miller et al., 2018). These datasets validate MAAT's ability to handle heterogeneous temporal dependencies (short-term sequences, long-term sequences, and mixed sequences) and anomaly types (point, contextual, collective) across domains ranging from aerospace to critical infrastructure. Their adoption in prior work (Xu et al., 2022; Yang et al., 2023b) ensures reproducibility and fair comparison. Table A.7 includes more details about the datasets.

4.2. Implementation

Our study follows the Anomaly Transformer model's protocol to evaluate our approach. We generate a sub-series using a non-overlapping sliding window. Anomaly detection involves scoring time points and setting a threshold based on the Anomaly Ratio detailed in the table. Other parameters, including batch size and model dimensionality, are also specified in Table B.8, ensuring robust evaluation and reproducibility of our improvements.

4.3. Evaluation metrics

The evaluation of anomaly detection in time series data employs various metrics to assess performance and capture the temporal continuity of anomalies. Point-based metrics, including Precision, Recall, and F1 Score, utilize true positives (TP), false positives (FP), and false negatives (FN) for evaluation Chandola et al. (2009). Affiliation metrics, such as Affiliation Precision (Aff-P) and Affiliation Recall (Aff-R), measure the fraction of predicted anomalies within true ranges and the fraction of true anomalies detected, respectively. Range-based metrics like Range-based Anomaly Recall (R_A_R) and Range-based Anomaly Precision (R_A_P) evaluate detection based on the overlap between true and predicted anomaly ranges. Additionally, volume-based metrics, including Volume-based ROC (V_ROC) and Volume-based Precision-Recall (V_PR), adjust true positive rate (TPR) and false positive rate (FPR) by incorporating the volume of anomalies and provide volume-weighted precision and recall. In summary, while point-based metrics offer basic performance insights, affiliation, range-based, and volume-based metrics deliver a more comprehensive evaluation by factoring in the temporal structure and duration of anomalies. In summary, while point-based metrics offer a basic assessment, affiliation, range-based, and volume-based metrics ((C.1), (C.2), (C.3), (C.4)) provide a comprehensive evaluation by accounting for temporal structure and anomaly duration (Hundman et al., 2018; Blazquez-Garcia et al., 2019).

5. Results

In this section, we introduce the results comparison in different test setups to showcase the effectiveness of the MAAT model.

5.1. Baseline results

After conducting various experiments, we present our results evaluated on precision Recall and F1 score metrics in Table 1: Our model Achieves remarkable improvements over existing methods, particularly when compared to the DCdetector and Anomaly Transformer, two leading approaches in the field. For instance,

On the **SMD** dataset, MAAT achieves a +**2.18%** F1-score and +**3.84%** Recall improvement over the **Anomaly Transformer**, attributable to its **Sparse Attention** mechanism, which prioritizes localized dependencies over redundant global interactions, refining anomaly localization. The +**8.64%** F1-score and +**13.93%** Recall gains over **DCdetector** stem from MAAT's **Mamba-SSM**, which preserves transient anomalies via skip connections while modeling long-term trends, circumventing DCdetector's contrastive loss limitations in retaining short-term deviations. Precision improvements (+**0.63%** and +**3.74%**) reflect **Gated Attention Fusion**'s adaptive suppression of false positives through context-aware reconstruction weighting. Compared to **LGAT**, MAAT's +**5.22%** F1, +**9.91%** Recall, and +**0.76%** Precision gains highlight its efficacy in isolating localized anomalies from periodic noise, where **LGAT's** graph attention struggles to resolve short-term signals amid relational smoothing.

On the **MSL** dataset, MAAT achieves a +**1.19%** F1-score and +**2.40%** Recall improvement over the **Anomaly Transformer**, attributable to its **Sparse Attention** mechanism, which prioritizes localized temporal dependencies (e.g., rapid actuator spikes) while suppressing ambient sensor noise. The +**0.32%** F1-score and +**0.95%** Recall gains over **DCdetector** arise from MAAT's **Mamba-SSM**, which models long-term degradation patterns while retaining short-term anomalies (e.g., actuator jitters) via skip connections, addressing DCdetector's limitations in preserving fine-grained deviations during contrastive learning. The marginal **–0.21%** Precision trade-off reflects MAAT's emphasis on recall for safety-critical aerospace applications, where missing anomalies poses greater risks than false alarms. Compared to **LGAT**, MAAT's +**4.19%** F1, +**8.20%** Recall, and +**0.08%** Precision improvements highlight its efficacy in detecting gradual wear-and-tear anomalies (e.g., actuator fatigue), where **LGAT's** graph attention over-smooths low-frequency signals through relational aggregation. These advancements are enabled by **Gated Attention Fusion**, which contextually balances abrupt fault detection against mission-phase baselines, ensuring robust anomaly discrimination in dynamic aerospace environments.

On the **SMAP** dataset, MAAT achieves a +**0.60%** F1-score and +**1.24%** Precision improvement over the **Anomaly Transformer**, driven by its **block-wise Sparse Attention**, which isolates abrupt anomalies (e.g., sensor spikes) while attenuating gradual seasonal trends, enhancing discrimination between transient faults and benign cycles. The +**0.93%** F1-score and +**1.39%** Recall gains over **DCdetector** stem from MAAT's **Mamba-SSM**, which models long-term cyclical patterns while preserving subtle anomalies (e.g., conductivity lags) via skip connections, addressing DCdetector's tendency to over-smooth short-term deviations during contrastive learning. The marginal **–0.08%** Recall trade-off against the Anomaly Transformer reflects MAAT's emphasis on precision for high-stakes telemetry validation, where false alarms are costlier than missed detections. Compared to **LGAT**, MAAT's +**3.69%** F1, +**7.30%** Recall, and +**0.14%** Precision improvements highlight its efficacy in detecting low-amplitude anomalies (e.g., soil moisture lags) within noisy seasonal data, where LGAT's graph attention over-smooths localized signals through relational aggregation. These advancements are enabled by **Gated Attention Fusion**, which balances sensitivity to rapid sensor anomalies and slow environmental shifts, leveraging context-aware reconstruction to minimize false positives—a critical capability for planetary science applications. On the **SWaT** dataset, MAAT achieves a +**0.09%** F1-score improvement over the **Anomaly Transformer** and +**0.08%** over **DCdetector**, with its **Sparse Attention** mechanism isolating transient cyber–physical attack signatures (e.g., valve tampering bursts) while attenuating high-frequency sensor noise, enhancing discrimination between malicious and benign industrial patterns. The **–0.28%** Precision trade-off against the Anomaly Transformer reflects MAAT's prioritization of recall (+**0.59%**) for safety-critical systems, where undetected attacks pose severe risks. The +**0.23%** Precision and +**0.04%** Recall gains over **DCdetector** highlight MAAT's **skip-connected Mamba-SSM**, which preserves transient anomalies (e.g., pump pressure drops) during long-term state–space denoising, addressing DCdetector's tendency to over-smooth short-lived deviations in contrastive learning. Against **LGAT**, MAAT's **–0.11%** F1 (offset by +**0.72%** Recall) underscores its focus on stealthy attack detection in 1-second industrial sequences, where LGAT's graph attention over-smooths temporal granularity through relational aggregation. These trade-offs are governed by **Gated Attention Fusion**, which balances anomaly sensitivity and noise suppression via context-aware reconstruction, ensuring robust detection in high-velocity, noisy cyber–physical environments.

On the **PSM** dataset, MAAT achieves a +**0.87%** F1-score and +**1.39%** Recall improvement over the **Anomaly Transformer**, driven by its **block-wise Sparse Attention**, which prioritizes localized temporal anomalies (e.g., CPU/memory spikes) while attenuating periodic workload noise, enhancing discrimination between transient anomalies and routine traffic patterns. The +**0.50%** F1-score and +**0.73%** Recall gains over **DCdetector** arise from MAAT's **Mamba-SSM**, which models long-term server metric consistency while retaining short-term disruptions (e.g., memory leaks) via skip connections, addressing DCdetector's over-smoothing of volatile sequences during contrastive learning.

Table 1
Comparison of various anomaly detection metrics across different datasets.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LOF	56.34	39.86	46.68	47.72	85.25	61.18	58.93	56.33	57.60	72.15	65.43	68.62	57.89	90.49	70.61
OCSVM	44.34	76.72	56.19	59.78	86.87	70.82	53.85	59.07	56.34	45.39	49.22	47.23	62.75	80.89	70.67
U-Time	65.95	74.75	70.07	57.20	71.66	63.62	49.71	56.18	52.75	46.20	87.94	60.58	82.85	79.34	81.06
Forrest	42.31	73.29	53.75	53.94	82.98	65.42	52.39	55.53	53.89	49.22	44.95	47.02	76.09	92.45	83.46
DAGMM	67.30	49.89	57.30	89.60	63.93	74.62	86.45	56.73	68.51	89.92	57.84	70.40	93.49	70.03	80.08
ITAD	86.22	73.71	79.48	69.44	84.09	76.07	82.42	66.89	73.85	63.13	52.08	57.08	72.80	64.02	68.13
VAR	78.35	70.26	74.08	71.68	81.42	77.90	81.38	53.88	64.83	81.59	60.29	69.34	90.71	83.82	87.13
MMPCACD	71.20	79.28	75.02	81.42	61.31	69.95	83.22	68.23	74.73	82.52	68.23	74.73	76.26	78.35	77.29
CL-MPPCA	82.36	76.07	79.09	73.71	88.54	80.44	63.16	72.88	67.72	76.78	81.50	79.07	56.02	**99.93**	71.80
TS-CP2	87.42	66.25	75.38	68.45	68.48	68.42	86.75	83.18	84.95	81.23	74.10	77.50	82.67	78.16	80.35
Deep-SVDD	78.54	79.67	79.10	**92.25**	76.63	83.26	56.02	69.04	62.40	80.42	84.45	82.39	95.41	86.79	90.73
BOCPD	70.90	82.04	76.07	80.32	87.60	83.62	86.45	85.85	86.14	84.96	70.75	79.01	82.72	75.33	77.70
LSTM-VAE	75.76	90.08	82.30	85.49	79.94	82.62	92.20	67.75	78.10	76.00	89.50	82.20	73.62	89.92	80.94
BeatGAN	78.55	88.92	83.42	85.42	87.88	86.64	92.38	55.85	69.61	76.04	87.46	81.32	90.30	93.84	92.04
LSTM	78.55	88.25	83.08	85.98	85.42	85.70	91.00	81.89	86.21	78.13	83.39	80.69	76.93	89.64	82.64
OmniAnomaly	83.68	88.52	85.22	90.14	89.50	89.82	81.42	84.30	82.83	81.42	84.30	82.83	83.38	74.46	78.64
THOC	79.65	91.10	85.02	88.45	90.97	89.69	92.06	89.34	90.68	92.48	98.32	95.33	97.14	98.74	97.94
AnomalyTrans	88.47	92.28	90.33	91.02	96.03	93.93	93.59	**99.41**	96.41	**93.59**	99.41	96.41	97.14	97.81	97.47
DCdetector	85.82	84.10	84.95	91.25	97.40	94.75	94.29	97.97	96.10	93.12	99.96	96.42	97.22	98.45	97.83
LGAT	88.27	85.91	87.08	91.38	90.13	90.86	94.61	92.03	93.30	93.28	99.91	96.48	**98.41**	98.25	**98.33**
Ours	**89.03**	**95.82**	**92.30**	**92.06**	**98.33**	**95.05**	**94.75**	99.33	**96.99**	93.33	**100.00**	**96.50**	97.48	**99.17**	98.32

Table 2
Comparison of Different Methods on NIPS-TS-GECCO and NIPS-TS-SWAN Datasets.

Dataset	NIPS-TS-GECCO			NIPS-TS-SWAN		
Metric	P	R	F1	P	R	F1
MatrixProfile (Benschoten et al., 2020)	4.6	18.5	7.4	17.1	17.1	17.1
GBRT (Taieb et al., 2012)	17.5	14.0	15.6	44.7	37.5	40.8
LSTM-RNN	17.0	22.6	19.3	45.2	35.8	40.0
Autoregression (Box et al., 2015)	39.2	31.4	34.9	42.1	35.4	38.5
OCSVM	18.5	**74.3**	29.6	47.4	49.8	48.5
IForest (Liu et al., 2008)	**43.9**	35.3	39.1	56.9	59.8	58.3
AutoEncoder	42.4	34.0	37.7	47.0	52.2	50.9
AnomalyTrans	25.7	28.5	27.0	90.7	47.4	62.3
DCdetector	38.3	59.7	46.6	95.5	59.6	73.4
Ours	42.4	70.0	**52.8**	**95.9**	**59.9**	**73.8**

The **+0.35%** Precision improvement over Anomaly Transformer and **+0.27%** over DCdetector reflects **Gated Attention Fusion**'s context-aware suppression of false alarms in high-density anomaly environments. Compared to **LGAT**, MAAT's near-parity F1-score (**–0.01%**) with a **+0.20%** Recall gain—despite a **–0.24%** Precision trade-off—demonstrates its focus on anomaly detection completeness in ambiguous scenarios (e.g., intermittent API failures), where LGAT's graph attention over-smooths short-term deviations through relational aggregation. This balance is enabled by MAAT's adaptive weighting of localized anomalies against reconstructed operational baselines, ensuring reliability in server monitoring without inflating false positives.

The slight variations we observe likely stem from DCdetector's dual-attention contrastive learning method, which works particularly well with highly variable datasets. For other datasets, however, we see somewhat reduced precision. These performance trade-offs highlight areas where MAAT could be improved, especially in contexts where avoiding false positives takes precedence over capturing all anomalies.

In Table 2,Our model establishes a new benchmark for anomaly detection on the NIPS-TS-GECCO and NIPS-TS-SWAN datasets, surpassing the best-performing methods in most metrics while demonstrating exceptional robustness across diverse challenges.

On the **NIPS-TS-GECCO** dataset, MAAT achieves a **6.2%** F1-score improvement over **DCdetector**, alongside **13.7%** and **23.2%** gains against **IForest** and **OCSVM**, respectively. These advancements stem from MAAT's **Sparse Attention** mechanism, which isolates transient water quality anomalies (e.g., pH spikes) from IoT sensor noise, addressing DCdetector's contrastive learning limitations in low-signal environments. The **16.7%** precision improvement over the **Anomaly Transformer** and **4.1%** over **DCdetector** reflects MAAT's **Mamba-SSM**, which suppresses false alarms by modeling long-term contamination trends while retaining abrupt deviations via skip connections. Though **IForest** achieves marginally higher precision (**–1.5%**), MAAT's recall-centric design prioritizes detection completeness in safety-critical water quality monitoring, where missed anomalies risk public health. On the **NIPS-TS-SWAN** dataset, MAAT delivers exceptional performance, surpassing the **Anomaly Transformer** by **5.2%** precision, **12.5%** recall, and **11.5%** F1-score through its ability to detect subtle solar magnetic anomalies (e.g., polarity misalignments) often obscured by spatially correlated noise. Against **DCdetector**, MAAT's **0.4%** precision, **0.3%** recall, and **0.4%** F1-score gains highlight its superior handling of long-term solar cycles (e.g., 11-year sunspot trends) via **Mamba-SSM**'s state–space modeling, which avoids DCdetector's over-smoothing during contrastive alignment. Traditional methods like **IForest** and **OCSVM** lag significantly, with MAAT achieving **39.0%** higher precision and **15.5%** F1-score over IForest, and **48.5%** precision, **25.3%** F1-score, and **10.1%** recall improvements over OCSVM. These results underscore MAAT's **Sparse Attention** efficacy in filtering solar noise while preserving localized magnetic flux cues, paired with **Mamba-SSM**'s capacity to model gradual solar phenomena without signal dilution. While minor trade-offs exist — such as a **–0.24%** precision reduction on **NIPS-TS-GECCO** and a marginal **–0.08%** recall dip on **NIPS-TS-SWAN** — MAAT's architecture prioritizes domain-specific robustness. For solar weather monitoring (SWAN), recall-driven anomaly detection mitigates catastrophic risks of undetected solar flares, whereas water quality applications (GECCO) balance precision to reduce false alarms. This adaptability, rooted in **Gated Attention**

Fusion's context-aware scoring, ensures MAAT excels in both high-dimensional IoT sensor streams and spatially correlated solar data, solidifying its versatility across diverse real-world anomaly detection.

5.2. In-depth performance analysis of MAAT vs. State-of-the-art models

Analysis of Performance Across Datasets Refer to Table 3 The performance of our model is comprehensively evaluated against AnomalyTrans and DCdetector, two state-of-the-art methods, across five benchmark datasets: MSL, SMAP, SWaT, PSM, and SMD. The results in Table 3 demonstrate that our model consistently outperforms or closely matches the best-performing methods across most metrics, establishing its superiority in anomaly detection tasks.

On the MSL dataset, our model surpasses AnomalyTrans in several metrics, including Accuracy Acc, where it achieves 98.92%, compared to AnomalyTrans's 98.69%. While DCdetector excels in precision-related metrics such as Aff-P 51.84% and R_A_P 91.64%, our model demonstrates robust recall metrics, achieving Aff-R 96.49% and R_A_R 90.97%, which are critical for detecting subtle anomalies in noisy environments.

On the **SMAP** dataset, MAAT exhibits marginally lower performance in volume-based metrics (V_ROC: 92.06%, V_PR: 90.70%) compared to AnomalyTrans (V_ROC: 95.52%, V_PR: 93.77%), reflecting inherent architectural trade-offs tailored for noisy, high-dimensional datasets. The fixed block-wise sparse attention mechanism, designed to suppress spurious correlations in noisy environments, prioritizes localized anomaly detection—enhancing robustness but limiting sensitivity to SMAP's weak long-range telemetry dependencies critical for volume-based aggregation. Concurrently, the gating mechanism optimizes denoising by favoring Mamba-SSM's state–space modeling on low-noise signals, improving global consistency at the cost of reduced contrast in fine-grained residuals. Furthermore, the association-discrepancy scoring emphasizes alignment between reconstruction errors and contextual associations, which prioritizes clear anomalies over subtle deviations in SMAP's gradually varying signals. While these design choices explain the modest gaps in V_ROC/V_PR, MAAT maintains strong recall (Aff-R: 94.27%), underscoring its reliability in detecting anomalies even in smooth telemetry. The trade-offs highlight MAAT's specialization: its architecture excels in bursty, noisy domains (e.g., industrial sensors, server clusters) but requires targeted adaptations for applications dominated by long-range, low-noise signals like SMAP.

On the SWaT dataset, our model achieves the highest Aff-P 56.26%, surpassing both AnomalyTrans 53.03% and DCdetector 52.40%. Additionally, our model demonstrates superior recall metrics, achieving Aff-R 97.79% and R_A_R 98.14%, which are marginally higher than DCdetector 97.67% and 98.43%, respectively. Our V _ROC 98.17% and V_PR 95.02% also outperform both competitors, highlighting its robustness in handling multivariate interactions and complex temporal patterns.

Our model achieves the best performance across nearly all metrics on the PSM dataset, a 1.5% increase in Aff-P vs. DCdetector, our model excels in recall-related metrics, achieving Aff-R 85.06%, R_A_R 94.15%, and R A_P 95.09%, outperforming AnomalyTrans and DCdetector. These results underscore our model's ability to handle high-dimensional sensor data with significant noise.

On the SMD dataset, our model achieves the highest scores across all metrics, setting a new benchmark for performance. It surpasses AnomalyTrans in 5.8% and 8.9% increase in F1-score and Aff-P respectively. Additionally, our model demonstrates superior recall metrics, achieving Aff-R 93.51%, R A_R 79.11%, and R_ A_P 75.41%, significantly outperforming both competitors. These gains highlight our model's effectiveness in capturing intricate temporal patterns in highly non-stationary multivariate time series.

In this section, we analyze MAAT's performance on the **NIPS-TS-SWAN** and **NIPS-TS-GECCO** datasets, as detailed in Table 4. On **NIPS-TS-SWAN**, MAAT achieves **95.9%** precision, **59.9%** recall, and **73.8%** F1-score, surpassing the **Anomaly Transformer** by +**5.2%** precision, +**12.5%** recall, and +**11.5%** F1, while outperforming **DCdetector** by +**0.4%** precision, +**0.3%** recall, and +**0.4%** F1. These gains stem from MAAT's **Sparse Attention** mechanism, which isolates subtle solar magnetic anomalies (e.g., polarity misalignments) from spatially correlated noise, and **Mamba-SSM**, which models long-term solar cycles (e.g., 11-year sunspot trends) without diluting transient flare precursors. Against traditional methods, MAAT dominates with +**39.0%** precision and +**15.5%** F1 over **IForest**, and +**48.5%** precision and +**25.3%** F1 over **OCSVM**, resolving their inability to handle high-dimensional, temporally complex solar data. On **NIPS-TS-GECCO**, MAAT achieves **42.4%** precision, **70.0%** recall, and **52.8%** F1, exceeding **DCdetector** by +**4.1%** precision, +**10.3%** recall, and +**6.2%** F1, and **Anomaly Transformer** by +**16.7%** precision, +**41.5%** recall, and +**23.7%** F1. The **Sparse Attention** mechanism excels here by filtering IoT sensor noise to detect transient water quality anomalies (e.g., pH spikes), while **Mamba-SSM** suppresses false alarms by modeling long-term contamination trends. Despite a marginal **–1.5%** precision trade-off against **IForest**, MAAT's +**10.3%** recall gain ensures comprehensive detection in safety-critical water quality monitoring, where missed anomalies risk public health. The affinity metrics further validate MAAT's robustness: on **NIPS-TS-GECCO**, **Aff-P** improves by +**7.88%** and **Aff-R** by +**11.89%** over **Anomaly Transformer**, demonstrating its ability to detect anomalies within true event ranges. The validation ROC corroborates this, with gains of +**1.55%** (SWAN) and +**11.42%** (GECCO), reflecting MAAT's balanced sensitivity-specificity trade-offs. These results underscore MAAT's architectural superiority: **Gated Attention Fusion** dynamically prioritizes recall for solar flare detection (SWAN) and precision for water quality (GECCO), while its **association discrepancy mechanism** aligns anomaly scores with contextual fidelity. By integrating localized anomaly isolation, long-term dependency modeling, and context-aware scoring, MAAT sets a new standard for time-series anomaly detection, delivering state-of-the-art performance across domains with divergent requirements—from solar magnetic analysis to IoT-driven environmental monitoring.

6. Discussion

6.1. Ablation study analysis

This section, presents the results of the ablation study of our model, emphasizing the significance of each component within MAAT and their effects on the baseline Anomaly Transformer model. The findings from the ablation study are summarized in Tables 5 and 6. Across five datasets SMD, MSL, SMAP, SWaT, and PSM the incorporation of Sparse Attention (SA), Mamba-SSM, and Gated Attention consistently shows enhancements over the baseline.

The baseline Anomaly Transformer demonstrates strong recall and moderate precision on **SMD** dataset. However, its dense self-attention mechanism makes it vulnerable to false positives due to noise. The addition of a Gating mechanism to the baseline (AnomalyTrans+Gating) improves both precision to 89.11% and recall to 93.09%, resulting in an F1-score increase to 91.05% - demonstrating that adaptive feature fusion alone provides substantial benefits. The integration of Sparse Attention (AnomTr+SA) maintains baseline precision but leads to a recall decrease of approximately 2.4%, resulting in an F1-score reduction of about 1.15%. This suggests that the block-wise formulation may oversimplify inter-variable dependencies and overlook subtle anomalies. The standalone Mamba model shows promising capabilities, maintaining a reasonable precision of 88.33% while improving recall to 93.53% (an increase of approximately 1.25% over the baseline), resulting in a modest F1-score gain to 90.86%. However, when combining Mamba with Sparse Attention without gating (Mamba+SA), we observe performance degradation with precision dropping to 87.28% and recall falling significantly to 86.76%, resulting in the lowest F1-score of 87.02%. This indicates that naively combining these approaches

Table 3
Performance comparison of AnomalyTrans, DCdetector, and our method across different datasets. Best results in **bold**, second-best underlined.

Dataset	Method	Acc	F1	Aff-P	Aff-R	R_A_R	R_A_P	V_ROC	V_PR
MSL	AnomalyTrans	98.69	93.93	51.76	95.98	90.04	87.87	88.20	86.26
	DCdetector	**99.06**	**96.60**	**51.84**	**97.39**	**93.17**	**91.64**	**95.15**	**91.66**
	Ours	98.92	95.05	51.58	96.49	90.97	88.72	90.66	88.49
SMAP	AnomalyTrans	99.05	96.41	51.39	**98.68**	**96.32**	94.07	**95.52**	**93.77**
	DCdetector	**99.21**	**97.02**	**51.46**	98.64	96.03	**94.18**	95.13	93.46
	Ours	99.03	96.29	49.34	94.27	93.59	92.02	92.06	90.70
SWaT	AnomalyTrans	98.51	94.22	53.03	90.88	97.73	96.32	97.99	94.39
	DCdetector	**99.09**	**96.33**	52.40	97.67	**98.43**	**96.96**	96.95	94.34
	Ours	98.97	95.93	**56.26**	**97.79**	98.14	95.00	**98.17**	**95.02**
PSM	AnomalyTrans	98.68	97.37	55.35	80.85	91.68	93.00	88.71	90.71
	DCdetector	98.95	97.94	54.71	82.93	91.55	92.93	88.41	90.58
	Ours	**99.06**	**98.32**	**55.53**	**85.06**	**94.15**	**95.09**	**90.77**	**92.67**
SMD	AnomalyTrans	98.75	87.18	54.36	90.12	74.95	73.00	78.71	70.71
	DCdetector	98.75	84.95	51.36	88.92	70.19	65.49	68.19	63.57
	Ours	**99.34**	**92.30**	**59.19**	**93.51**	**79.11**	**75.41**	**79.09**	**75.41**

Table 4
Multi-metric results on NIPS-TS datasets. All results in %. Best results in **bold**, second-best underlined.

Dataset	Method	Acc	P	R	F1	Aff-P	Aff-R	R_A_R	R_A_P	V_ROC
NIPS-TS-SWAN	AnomalyTrans	84.57	90.71	47.43	62.29	**58.45**	**9.49**	86.42	93.26	84.81
	DCdetector	85.94	95.48	59.55	73.35	50.48	5.63	88.06	94.71	86.25
	Ours	**86.10**	**95.93**	**59.91**	**73.76**	58.22	6.72	**88.15**	**94.85**	**86.36**
NIPS-TS-GECCO	AnomalyTrans	98.03	25.65	28.48	29.09	49.23	81.20	56.35	22.53	55.45
	DCdetector	98.56	38.25	59.73	46.63	50.05	88.55	62.95	34.17	62.41
	Ours	**98.68**	**42.41**	**70.00**	**52.82**	**57.11**	**93.09**	**64.96**	**38.01**	**66.87**

without proper integration mechanisms is suboptimal. Remarkably, the complete MAAT model, which combines Sparse Attention and Mamba-SSM through a dynamic Gated Attention mechanism (Eqs. (8)–(9)), achieves a net F1 improvement of nearly 2% over the baseline, fueled by a modest precision gain of around 0.6% and a significant recall enhancement of roughly 3.5%. This adaptive gating effectively balances the contributions of precision and long-term pattern refinement, robustly mitigating noise while preserving critical anomaly signals within the complex SMD environment.

For the **MSL** dataset, which comprises sensor and actuator data characterized by high variability and transient noise, the baseline Anomaly Transformer demonstrates strong recall but moderate precision, resulting in an F1-score of approximately 93.93%. However, it encounters challenges with false positives induced by noise. The addition of Gating alone (AnomalyTrans+Gating) yields notable improvements, with precision increasing to 91.81% and recall to 97.59%, resulting in an F1-score of 94.61% - a significant 0.68% improvement over the baseline. The introduction of Sparse Attention (AnomTr+SA) alleviates noise issues, yielding a precision improvement of around 0.7%, albeit at the expense of a 2% reduction in recall, which ultimately lowers the F1-score by about 1.1%. This trade-off occurs because Sparse Attention effectively filters out noise but inadvertently excludes crucial short-term patterns. The standalone Mamba model demonstrates strong capabilities with a precision of 92.06% and recall of 97.59%, achieving an F1-score of 94.74%. When combined with Sparse Attention (Mamba+SA), we observe further improvement in recall to 98.07% while maintaining the same precision, resulting in an enhanced F1-score of 94.97%. This suggests that Mamba effectively captures long-range dependencies while Sparse Attention helps focus on locally relevant patterns. The integration of Mamba-SSM with the Anomaly Transformer (AnomTr+mamba) enhances precision to 92.17% while maintaining competitive recall at 96.00%, resulting in a solid F1 improvement to 94.05%. The comprehensive MAAT model synthesizes these advancements through Gated Attention, achieving the highest F1-score of 95.05% - a gain of 1.12% over the baseline, driven by maintaining strong precision at 92.06% and achieving the highest recall of 98.33%. During instances of transient noise, the adaptive gating mechanism (Eq. (8)) emphasizes the denoised outputs of Mamba while preserving the localized anomaly signals from Sparse Attention. This balanced integration strengthens MAAT's robustness against short-lived disturbances, significantly enhancing anomaly detection in highly dynamic environments.

The **SMAP** dataset, which comprises satellite telemetry data characterized by long-range dependencies and subtle anomalies, underscores MAAT's ability to balance global context with localized anomaly detection. The Anomaly Transformer achieves an almost perfect recall of 99.41% but suffers from overfitting, which limits its precision to 93.59%, resulting in an F1-score of 96.41%. The addition of Gating alone (AnomalyTrans+Gating) maintains similar performance with precision at 93.60%, recall at 99.35%, and F1-score at 96.39%, suggesting that adaptive feature fusion provides limited benefits on this dataset. To mitigate overfitting, Sparse Attention addresses redundant global interactions (see Eq. (3)), although this leads to slightly weaker recall of 99.14% and a minor decrease in F1-score to 96.26%. The standalone Mamba model achieves comparable results to the baseline with precision of 93.57%, recall of 99.30%, and F1-score of 96.35%. When combined with Sparse Attention (Mamba+SA), performance remains stable with precision at 93.58%, recall at 99.36%, and F1-score at 96.39%. Notably, the combination of Anomaly Transformer with Mamba (AnomTr+mamba) yields the most significant improvement, with precision substantially increasing to 95.59% while maintaining a competitive recall of 98.62%, resulting in the highest F1-score of 97.08%. This suggests that incorporating state space modeling into the Transformer architecture effectively balances precision and recall for this dataset. The complete MAAT model harmonizes these strengths through Gated Attention (refer to Eqs. (8)–(9)), achieving a final precision of 94.75%, a recall of 99.33%, and an F1-score of 96.99%, exceeding the baseline by 0.58% but slightly underperforming compared to AnomTr+mamba. This enhancement is propelled by the gating mechanism, which prioritizes Mamba's global context for slow-evolving anomalies while leveraging Sparse Attention's local focus for abrupt changes. This adaptability enables MAAT to effectively capture both transient and long-term anomalies, making it well-suited to address the unique challenges presented by the SMAP dataset, though further optimization of the Mamba integration could potentially yield even better results.

The **SWaT** dataset, which includes 51-dimensional sensor data from a complex water treatment system, highlights the importance of capturing inter-sensor dependencies for effective anomaly detection. The baseline Anomaly Transformer achieves an impressive recall of 99.41%; however, it faces challenges with false positives, resulting in a precision of 93.59% and an F1-score of 96.41%. The addition of Gating alone (AnomalyTrans+Gating) significantly improves recall to a perfect 100% but substantially reduces precision to 83.53%, causing the F1-score to drop to 91.03%. This suggests that while gating helps capture more anomalies, it struggles with discrimination in this complex multi-sensor environment. The introduction of Sparse Attention (AnomTr+SA) also reduces precision to 89.53% and slightly decreases recall to 98.86%, resulting in an F1-score drop to 93.96%, as this approach fails to model critical cross-sensor relationships adequately. The standalone Mamba model demonstrates remarkable capabilities, maintaining a strong precision of 92.77% while achieving perfect recall of 100%, resulting in an impressive F1-score of 95.98%. When combined with Sparse Attention (Mamba+SA), performance further improves with precision increasing to 93.26% while maintaining perfect recall, yielding an enhanced F1-score of 96.51% that surpasses the baseline. This suggests that the state space model effectively captures the complex temporal dependencies in the water treatment system while sparse attention helps focus on locally relevant patterns. The integration of Mamba-SSM with the Anomaly Transformer (AnomTr+mamba) achieves the best performance with precision at 93.33%, perfect recall at 100%, and the highest F1-score of 96.55%, by effectively capturing complex interdependencies while filtering out redundant sensor noise. The complete MAAT model strategically combines these mechanisms through Gated Attention, achieving identical precision and recall values, with a marginally lower F1-score of 96.50%. This development underscores Mamba's ability to recognize the limitations of sparse attention, ensuring that MAAT maintains a balanced approach to precision and recall in a multivariate anomaly detection context, though the simpler AnomTr+mamba variant performs slightly better on this specific dataset.

The **PSM** dataset, consists of 25-dimensional sensor data characterized by substantial noise, underscores the importance of reducing false positives while ensuring effective anomaly detection. The baseline Anomaly Transformer exhibits commendable performance, attaining an F1-score of 97.47% with precision of 97.14% and recall of 97.81%. However, it suffers from noise-induced misclassifications in this high-dimensional environment. The addition of Gating alone (AnomalyTrans+Gating) maintains similar performance with minimal changes, achieving precision of 97.10%, recall of 97.61%, and an F1-score of 97.35%, suggesting that adaptive feature fusion provides limited benefits for this dataset. In contrast, Sparse Attention (AnomTr+SA) maintains competitive precision at 97.27% while significantly improving recall to 99.37%, consequently elevating the F1-score to 98.31%. This enhancement is due to its localized attention windows (Eq. (3)), which effectively focus on critical time steps in this noisy environment. The standalone Mamba model demonstrates strong performance with precision of 97.10% and substantially improved recall of 98.72%, resulting in an F1-score of 97.90%. When combined with Sparse Attention (Mamba+SA), precision increases to 97.41% with a slight reduction in recall to 98.37%, maintaining a similar F1-score of 97.89%. The integration of Mamba-SSM with the Anomaly Transformer (AnomTr+mamba) maintains competitive precision at 97.41% and recall at 98.37%, resulting in an F1-score of 97.89%. The comprehensive MAAT model combines these strengths through Gated Attention (Eqs. (8)–(9)), resulting in a balanced precision of 97.48%, an excellent recall of 99.17%, and the highest F1-score of 98.32%. This adaptability is essential in the noisy, high-dimensional context of PSM, where traditional attention mechanisms often struggle to identify true anomalies. Additionally, the Association Discrepancy metric (Eq. (10)) refines anomaly detection by penalizing deviations from expected sensor dependencies, thereby reducing false positives. These findings reinforce the effectiveness of MAAT in dynamically balancing precision, recall, and robustness across a range of time-series anomaly detection tasks, with the full model outperforming all individual components on this challenging dataset.

The **NIPS-TS-GECCO** dataset is a benchmark for drinking water quality, characterized by sporadic, high-frequency anomalies and noise from IoT sensors. This dataset highlights the limitations of the baseline Anomaly Transformer, which attains an F1-score of only 29.09%. The addition of Gating alone (AnomTr+Gating) dramatically reduces precision to just 14.01% while substantially improving recall to 54.25%, resulting in a decreased F1-score of 22.27%. This suggests that simple gating mechanisms struggle with discriminating true anomalies from noise in this challenging environment. The standalone Mamba model performs poorly on this dataset, with precision plummeting to 5.29% despite achieving the highest recall of 62.19%, resulting in the lowest F1-score of just 9.75%. However, when Mamba is combined with Sparse Attention (Mamba+SA), performance dramatically improves with precision rising to 37.06% and recall maintained at 51.78%, yielding a competitive F1-score of 43.20%. This demonstrates that Mamba requires focused attention mechanisms to be effective on datasets with sporadic anomalies. Implementing Sparse Attention alone (AnomTr+SA) yields the most significant improvements, with precision reaching 45.19% and an exceptional recall of 84.25%, resulting in the highest F1-score of 58.82%. These enhancements are achieved by focusing on critical time steps using Ω_i (Eq. (3)), effectively reducing the conflation of transient anomalies with sensor noise. Conversely, incorporating Mamba-SSM with the Anomaly Transformer (AnomTr+mamba) results in more modest gains, with precision of 36.00% and recall of 51.78%, yielding an F1-score of 42.47%. The linear dependency modeling is less effective in addressing abrupt, short-term anomalies that characterize this dataset. The comprehensive MAAT model balances these approaches, achieving a precision of 42.41% and recall of 70.00%, resulting in an F1-score of 52.82% - lower than AnomTr+SA but significantly higher than most other variants. The gating mechanism prioritizes Sparse Attention during sudden events ($g \to 0$) while leveraging Mamba's contextual modeling during stable periods ($g \to 1$). In this context, the full MAAT architecture may be overparameterized for the specific characteristics of the GECCO dataset, where the simpler AnomTr+SA approach proves more effective for detecting the sporadic, short-lived anomalies in drinking water quality data.

The **NIPS-TS-SWAN** dataset, derived from solar photospheric vector magnetograms, focuses on detecting subtle anomalies within high-dimensional and temporally complex patterns of solar activity. The baseline Anomaly Transformer (AnomalyTrans) demonstrates moderate performance, achieving a precision of 90.71%, recall of 47.43%, and an F1-score of 62.29%. However, its dense self-attention mechanism has difficulty distinguishing rare solar anomalies from noisy, non-stationary signals. The addition of Gating alone (AnomTr+Gating) substantially improves model performance, increasing precision to 96.73% and recall to 59.04%, resulting in a significant F1-score improvement to 73.33% - an 11.04% absolute gain over the baseline. This demonstrates that adaptive feature fusion is particularly effective for solar magnetogram analysis. The standalone Mamba model shows similar impressive performance with precision of 96.78%, recall of 59.30%, and an F1-score of 73.54%, highlighting the effectiveness of state space modeling for capturing complex solar dynamics. When combined with Sparse Attention (Mamba+SA), performance marginally improves with precision rising to 96.89% and recall to 59.37%, yielding an F1-score of 73.63%. By utilizing Sparse Attention alone (AnomTr+SA), the model achieves excellent precision of 97.03% and recall of 59.28%, resulting in an F1-score of 73.60% - representing a 6.32% absolute improvement in precision and 11.85% in recall compared to the baseline. This variant effectively focuses on localized time windows critical for transient solar events. The AnomTr+mamba combination yields the highest precision of 97.06% with competitive recall of 59.34%, resulting in an F1-score of 73.65% - demonstrating how Mamba enhances global context for solar

Table 5
Ablation study of Precision, Recall, and F-Score over five datasets. "SA" stands for Sparse Attention; the best metrics in each column are in **bold** and the second-best are underlined.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
AnomalyTrans	88.47	92.28	90.33	91.02	96.03	93.93	93.59	**99.41**	96.41	**93.59**	99.41	96.41	97.14	97.81	97.47
AnomalyTrans+Gating	**89.11**	93.09	91.05	91.81	97.59	94.61	93.60	99.35	96.39	83.53	**100.00**	91.03	97.10	97.61	97.35
Mamba	88.33	93.53	90.86	92.06	97.59	94.74	93.57	99.30	96.35	92.77	**100.00**	95.98	97.10	98.72	97.90
Mamba + SA	87.28	86.76	87.02	92.06	98.07	94.97	93.58	99.36	96.39	93.26	**100.00**	96.51	97.41	98.37	97.89
AnomTr+SA	88.47	89.89	89.18	91.67	94.05	92.84	93.55	99.14	96.26	89.53	98.86	93.96	97.27	**99.37**	98.31
AnomTr+mamba	87.52	90.04	88.76	**92.17**	96.00	94.05	**95.59**	98.62	**97.08**	93.33	**100.00**	**96.55**	97.41	98.37	97.89
Ours	89.03	**95.82**	**92.30**	92.06	**98.33**	**95.05**	94.75	99.33	96.99	93.33	**100.00**	96.50	**97.48**	99.17	**98.32**

pattern recognition. The full MAAT model integrates these components through Gated Attention, achieving a slightly lower precision of 95.93% but the highest recall of 59.91%, resulting in the best overall F1-score of 73.76%. This represents a 5.22% precision gain, 12.48% recall improvement, and 11.47% F1-score increase over the baseline. The gating mechanism adaptively balances attention, prioritizing Mamba's long-term dynamics during stable periods (when $g \to 1$) and emphasizing Sparse Attention's localized responses during rapid anomalies (when $g \to 0$). While the performance differences between the top-performing variants are small on this dataset, MAAT's balanced approach results in the most robust overall performance for detecting the complex patterns in solar activity data. The Association Discrepancy metric enhances anomaly detection by penalizing the differences between expected solar magnetic correlations and the observed sparse dependencies. This effectively identifies inconsistencies in the alignments of magnetic flux tubes. However, despite these improvements, MAAT's recall rate is still slightly lower than IForest's. This demonstrates a precision–recall trade-off, in which the model focuses on high-confidence anomalies rather than weaker signals. This is a critical consideration for real-world solar monitoring, where it is essential to minimize false positives.

Impact of block-wise sparse attention on anomaly transformer performance

To isolate the effect of replacing dense self-attention with block-wise sparse attention (SA), we compare the vanilla Anomaly Transformer (AT) against AT+SA across eight benchmark datasets. Table 5 reports Precision (P), Recall (R) and F1 for each variant; here we analyze the qualitative behavior by dataset:

SMD (Server-Machine Dataset). SMD comprises multivariate resource-usage traces from dozens of machines, where anomalies manifest as slow, coordinated degradations spanning hundreds of timesteps. AT+SA's strict locality truncates cross-series dependencies, dropping recall by 2.4 percent (92.28 to 89.89%) and net F1 by 1.15 percent. This decline underscores that purely local windows cannot capture the subtle, multi-stream fault signatures that emerge only in longer-range correlations.

MSL (Mars Science Laboratory). MSL anomalies arise from slow drifts in rover instrumentation, often correlated across sensor arrays. With SA, AT loses visibility into these distributed trends, reducing recall by 2.0 percent and F1 by 1.09 percent. The marginal precision gain (91.02 to 91.67%) reflects cleaner local reconstructions, but at the expense of missing larger drift patterns.

SMAP (Soil Moisture Active Passive). SMAP features mixed-scale anomalies: isolated spikes superimposed on seasonal drifts. AT+SA yields only a slight F1 drop –0.15 percent because AT's dense attention already excels at spike detection, and the residual drift component can still be partially captured via local context when drifts are slow relative to block size.

SWaT (Secure Water Treatment). SWaT events involve coordinated valve and sensor behavior over extended intervals. AT+SA severs these long-term dependencies, causing recall to fall from 99.41 to 98.86% and F1 to drop by 2.45 percent. The loss of cross-component context highlights the importance of global modeling for industrial control anomalies.

PSM (Pump Sensor Monitoring). PSM records a single-pump vibration signal with sharp, transient pressure spikes. Here, AT+SA's strict locality is an advantage: recall increases from 97.81 to 99.37% and F1 by +0.84 percent. The model can focus exclusively on immediate neighbors, yielding crisper detection of short-lived anomalies without dilution by distant noise.

SWAN and GECCO. These synthetic tasks combine periodic bursts with long-range dependencies up to 64 K steps. AT+SA produces modest F1 gains (+1.0 to +1.5 pts) by capturing the periodic components, but its inability to model the power-law drift components limits overall improvement.

Hybrid drift+noise (controlled synthetic series). In mixed scenarios of periodic spikes plus gradual drift, AT+SA reduces reconstruction error on the spike component, improving local recall by 1.2 pts, but fails to track the drift—manifesting as elevated reconstruction residuals during slow trend shifts.

Key Insight: Block-wise sparse attention sharpened local anomaly detection (e.g. PSM, synthetic bursts) by concentrating model capacity on immediate neighbors, but invariably degrades performance on multi-series or long-term drift anomalies (SMD, MSL, SWaT). This differential effect motivates the subsequent integration of a global, linear-time Mamba-SSM branch and adaptive gating to recover lost context while retaining SA's local precision.

6.2. Reconstruction loss

To rigorously evaluate the reconstruction performance of our method against the Anomaly Transformer baseline, we conducted a comparative analysis of their respective reconstruction losses, as illustrated in Fig. D.2. We computed the logarithmic difference between the two losses for each batch, defined as:

$$\Delta L_i = \log(L_i^{AT}) - \log(L_i^{MAAT}) \tag{11}$$

where L_i^{AT} and L_i^{MAAT} represent losses for each batch i. The results are visualized in Fig. D.2, with green columns indicating lower loss for our method ($\Delta L_i > 0$) and red columns showing a superior performance by AT ($\Delta L_i < 0$). This illustration highlights the advantages of our approach during training.

MAAT outperforms AT on the SMD dataset with an F1 score of 92.30%, benefiting from effective reconstruction loss minimization (Fig. D.2(a)). The differential analysis mainly shows positive ΔL_i, underscoring the effectiveness of Sparse Attention and Mamba-SSM in capturing critical time steps and long-range dependencies. Minor precision fluctuations do not compromise robustness, as evidenced by an Affiliation Recall of 93.51% and a Range-based Recall of 79%.

On the MSL dataset, MAAT also achieves lower reconstruction loss (Fig. D.2(a)), with a predominance of green bars in the ΔL_i plot, indicating its capacity to manage noisy patterns from Mars rover sensors. Despite slight precision trade-offs, MAAT shows significantly higher recall (96.03%) and robust Affiliation Recall (96.49%).

In the SMAP dataset, MAAT excels with an F1 score of 96.99%, consistently maintaining lower reconstruction loss (Fig. D.2(f)). The logarithmic difference reveals a dominance of green bars, highlighting

Table 6
Ablation study Precision, Recall, and F1-Score on NIPS datasets. "SA" stands for Sparse Attention; the best metrics are in **bold**.

Dataset	NIPS SWaN			NIPS GECCO		
Metric	P	R	F	P	R	F
AnomalyTrans	90.71	47.43	62.29	25.65	28.48	29.09
AnomTr+Gating	96.73	59.04	73.33	14.01	54.25	22.27
Mamba	96.78	59.30	73.54	05.29	62.19	09.75
Mamba + SA	96.89	59.37	73.63	37.06	51.78	43.20
AnomTr+SA	97.03	59.28	73.60	**45.19**	**84.25**	**58.82**
AnomTr+mamba	**97.06**	59.34	73.65	36.00	51.78	42.47
Ours	95.93	**59.91**	**73.76**	42.41	70.00	52.82

MAAT's proficiency in identifying subtle anomalies in satellite telemetry. Although there is a minor recall trade-off, precision gains enhance Affiliation Precision (49.34%).

Lastly, on the PSM dataset, MAAT achieves an F1 score of 98.32%, demonstrating its effectiveness with high-dimensional sensor data (Fig. D.2(c)). While facing occasional red bars in the ΔL_i plot, MAAT maintains superior Affiliation Recall (85.06%) and Range-based Precision (95.09%), ensuring accurate anomaly localization in noisy sequences.

7. Conclusions and future works

In this work, we introduced the Mamba Adaptive Anomaly Transformer (MAAT), which enhances unsupervised time series anomaly detection for practical applications like industrial monitoring and environmental sensing. MAAT effectively captures short- and long-term temporal dependencies by improving association discrepancy modeling with a new Sparse Attention mechanism and integrating a Mamba-Selective State Space Model (Mamba-SSM). Gated Attention mechanisms allow for a flexible combination of features from the original reconstruction and the Mamba-SSM output, striking a balance between accuracy and complexity. Moreover, MAAT trains efficiently, making it suitable for resource-limited settings. Evaluation of benchmark datasets has revealed that MAAT significantly surpasses alternative methods, including the Anomaly Transformer and DCdetector, in both anomaly detection accuracy and generalization across various time series. It mitigates challenges such as sensitivity to short context windows and computational inefficiencies that have been prevalent in prior methodologies. Despite its efficiency in training, MAAT still has limitations. It is susceptible to hyperparameters, especially when balancing the reconstruction module with the Mamba-SSM pathway. Tuning these parameters is essential for optimizing performance across various datasets. Additionally, although we have designed our method to reduce the effects of noise and non-stationary conditions, its performance may decline in situations with very high noise or when abnormal patterns closely mimic normal ones. Finally, while MAAT shows reduced computational demands compared to some deep learning options, further enhancement of its inference speed is necessary for real-time applications. There is potential for future research in several areas. We could explore adaptive, data-driven hyperparameter tuning to stabilize the model and boost its performance in various conditions. Incorporating online learning or incremental updates could enhance MAAT functionality in dynamic environments where quick anomaly detection is crucial. Additionally, exploring hybrid models that combine the strengths of reconstruction-based approaches with contrastive learning may yield better results, especially in challenging situations characterized by significant non-stationarity or noise. By pursuing these avenues, future work can build on MAAT and advance unsupervised time series anomaly detection in real-world applications.

CRediT authorship contribution statement

Abdellah Zakaria Sellam: Conceptualization, Writing – original draft, Methodology. **Ilyes Benaissa:** Software, Data curation. **Abdelmalik Taleb-Ahmed:** Writing – review & editing, Validation, Investigation. **Luigi Patrono:** Conceptualization. **Cosimo Distante:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Mr. Arturo Argentieri from CNR-ISASI Italy for his technical contribution to the multi-GPU computing facilities. This research was partially funded by the Italian Ministry of Health, Italian Health Opera National Plan (Cohesion and Development Fund 2014–2020), trajectory 2 "eHealth, advanced diagnostics, medical device and mini invasiveness", project "Sistema di Monitoraggio ed Analisi basato su intelligenza aRTificiale per pazienti affetti da scompenso CARdiaco cronico con dispositivi medici miniinvasivi e indossabili Evoluti – SMART CARE" (CUP F83C22001380006), by the European Union - Next Generation EU, PRIN 2022 PNRR call, under the project "Interactive digital twin solutions for cardiovascular disease Management, PReventiOn and treatment leVeraging the internet of things and Edge intelligence paradigms - IMPROVE, and funded in part by Future Artificial Intelligence Research—FAIR CUP B53C22003630006 grant number PE0000013.

Appendix A. Dataset

We evaluate our model on seven benchmark datasets derived from real-world applications, emphasizing diverse temporal dependencies and anomaly types. Each dataset is structured as follows:

We evaluate MAAT across seven real-world telemetry benchmarks. The MSL (Mars Science Laboratory) dataset comprises sensor and actuator readings from NASA's rover, featuring mixed temporal dynamics — short-term actuator sequences and temperature spikes alongside long-term wear trends and mission-phase cycles — and includes both point anomalies (sudden sensor failures) and contextual anomalies (unexpected actuator behavior) (NASA, 2018a). SMAP (Soil Moisture Active Passive) captures soil moisture telemetry from an Earth-observing satellite, with dominant long-term patterns such as seasonal moisture cycles and orbital shifts; its anomalies split roughly 62% point errors in telemetry and 38% contextual drifts across correlated sensors (NASA, 2018b). PSM (eBay Server Machines) provides 25-dimensional, second-level CPU, memory, and network I/O metrics with daily periodicity, where both point hardware spikes and contextual irregular resource usages occur (Huang et al., 2021). The SMD (Server Machine Dataset) logs minute-level CPU/memory and weekly workload cycles on 28 machines, exhibiting collective failures across nodes and contextual maintenance deviations (Su et al., 2019). SWaT (Secure Water Treatment) is a high-resolution industrial control set of 51 sensor streams (flow rates, chemical dosages, pressures) sampled every second, with point anomalies from tampered readings or cyber-attacks and cascading process failures as collective anomalies (Mathur and Tippenhauer, 2016). NIPS-TS-SWAN (Solar Weather) tracks solar magnetic dynamics hourly, plus minute-level flux sequences, revealing subtle contextual misalignments and collective flare precursors (NASA, 0000). Finally, NIPS-TS-GECCO (Drinking Water Quality) records minute-level pH,

Table A.7
Details of benchmark datasets. AR (anomaly ratio) represents the abnormal proportion of the whole dataset.

Benchmark	Temporal Scope	Dim.	#Training	#Test (Labeled)	AR (%)	Anomaly Types
MSL	Mixed	55	58,317	73,729	10.5	Contextual, Point
SMAP	Mixed (with dominant long-term patterns)	25	135,183	427,617	12.8	Point (62%), Contextual (38%)
PSM	Short-term	25	132,481	87,841	27.8	Collective, Point
SMD	Mixed	38	708,405	708,420	4.2	Contextual, Collective
SWaT	Short-term	51	495,000	449,919	12.1	Point (cyber-attacks)
NIPS-TS-GECCO	Short-term	9	69,260	69,261	1.1	Point
NIPS-TS-SWAN	Long-term	38	60,000	60,000	32.6	Collective, Contextual

Table B.8
Hyperparameters for MAAT model across datasets.

Dataset	Window Size	Batch Size	d_{model}	Anomaly Ratio (%)
SMD	100	128	512	0.5
MSL	100	256	512	0.85
SMAP	105	128	512	0.5
PSM	100	128	512	1.0
SWaT	100	256	512	0.5
NIPS-TS-GECCO	100	32	512	0.5
NIPS-TS-SWAN	100	128	512	0.9

turbidity, and chlorine readings over days, combining sporadic point malfunctions with longer-term contamination trends (Miller et al., 2018).

Table A.7 shows a summary of each benchmarking dataset.

Appendix B. Hyperparameters for model training

Table B.8 outlines the hyperparameters utilized for training the MAAT model across various datasets. The dimensions of the window, which set the length of the input sequence processed at each step, is set to 100 for most datasets. However, for SMAP, a slightly larger window of 105 is employed to better capture long-range dependencies. The batch size is tailored to the specific characteristics of each dataset and the computational constraints, with values ranging from 32 (for NIPS-TS-GECCO) to 256 (for MSL and SWaT). This range guarantees effective gradient updates while preserving training stability. The model dimension (d_{model}) remains consistently set at 512 for all datasets, defining the size of feature representations and ensuring consistency in learned embeddings. Additionally, the anomaly ratio, representing the proportion of data points identified as anomalies in each dataset, ranges from 0.5% to 1%, which affects the model's responsiveness to infrequent events. These hyperparameters have been carefully chosen to enhance performance across various anomaly detection contexts.

AR Threshold Selection

The threshold selection process follows these steps:

- We first calculate anomaly scores (Eq. (1)) for all points in the unlabeled validation subset after completing model training.
- Upon analyzing the frequency distribution of these scores, we observed that they naturally separate into two clusters: - A large cluster corresponding to normal data points with lower anomaly scores. - A smaller cluster representing potential anomalies with higher anomaly scores.
- This smaller cluster contains approximately r time points, where r converges to dataset-specific values: 0.1% for SWaT, 0.5% for SMD, and 1% for other datasets.
- Given that test data remains inaccessible in practical deployments, we establish a fixed threshold value δ that ensures all validation set points with anomaly scores exceeding δ (comprising proportion r of the data) are flagged as anomalies.

In operational settings, anomaly detection capacity is ultimately bounded by available human resources for investigation. Therefore, controlling the detection rate through ratio parameter r provides a more practical approach than absolute thresholds, allowing organizations to align anomaly detection volume with their investigative capabilities.

Appendix C. Evaluation metrics

Evaluating anomaly detection in time series data requires metrics point-wise performance and capturing anomalies' temporal continuity. This work uses traditional and specialized point-based metrics to evaluate contiguous anomaly segments.

C.0.1. Point-based metrics

The standard metrics include Precision, Recall, and F1 Score, calculated using the counts of true positives (TP), false positives (FP), and false negatives (FN). The F1 Score is the harmonic mean of Precision and Recall (Chandola et al., 2009).

C.0.2. Affiliation metrics: Aff-P and Aff-R

These metrics evaluate partial detection within multi-step anomaly segments.

Affiliation precision (Aff-P):. Measures the fraction of predicted anomalies within true anomaly ranges:

$$\text{Aff-P} = \frac{\sum_{t\in\mathcal{P}} \mathbf{1}\{t \in \mathcal{T}\}}{|\mathcal{P}|}, \tag{C.1}$$

where $\mathcal{P}$ is the predicted anomalies, $\mathcal{T}$ is the true anomalies, and $\mathbf{1}\{\cdot\}$ is the indicator function.

Affiliation recall (Aff-R):. Measures the fraction of true anomalies detected:

$$\text{Aff-R} = \frac{\sum_{t\in\mathcal{T}} \mathbf{1}\{t \in \mathcal{P}\}}{|\mathcal{T}|}. \tag{C.2}$$

C.0.3. Range-based metrics

These assess detection over entire anomaly segments.

Range-based anomaly recall (R_A_R):. A true anomaly range R_i^{true} is detected if overlap with any predicted range exceeds threshold τ:

$$\text{R\_A\_R} = \frac{\sum_{i=1}^{N_{\text{true}}} \mathbf{1}\left\{\max_j \text{Overlap}\left(R_i^{\text{true}}, R_j^{\text{pred}}\right) \geq \tau\right\}}{N_{\text{true}}}. \tag{C.3}$$

Range-based anomaly precision (R_A_P):.

$$\text{R\_A\_P} = \frac{\sum_{j=1}^{N_{\text{pred}}} \mathbf{1}\left\{\max_i \text{Overlap}\left(R_j^{\text{pred}}, R_i^{\text{true}}\right) \geq \tau\right\}}{N_{\text{pred}}}. \tag{C.4}$$

C.0.4. Volume-based metrics: V_ROC and V_PR

These incorporate anomaly duration into evaluation.

Table E.9
Comparison of model complexity across different time series anomaly detection models.

Model	Parameters	FLOPs
AnomalyTransformer	1.64M	0.33 GFLOPs
AnomalyTransformer+Sparse Attention	1.59M	0.31 GFLOPs
DCdetector	0.91M	3.58 GFLOPs
MAAT (Ours)	2.19M	0.41 GFLOPs

Volume-based ROC (V_ROC):. Adjusts TPR and FPR by anomaly volume:

$$\text{TPR}_{\text{vol}} = \frac{\text{Volume of Correct Detections}}{\text{Total Volume of True Anomalies}},$$
$$\text{FPR}_{\text{vol}} = \frac{\text{Volume of False Detections}}{\text{Total Volume of Normal Data}}.$$

The V_ROC curve plots TPR_{vol} vs. FPR_{vol}.

Volume-based precision-recall (V_PR):. Volume-weighted precision and recall are defined as follows:

$$\text{Precision}_{\text{vol}} = \frac{\text{Volume of Correct Detections}}{\text{Volume of All Detections}},$$
$$\text{Recall}_{\text{vol}} = \frac{\text{Volume of Correct Detections}}{\text{Total Volume of True Anomalies}}.$$

The V_PR curve plots $\text{Precision}_{\text{vol}}$ vs. $\text{Recall}_{\text{vol}}$.

In summary, while point-based metrics offer a basic assessment, affiliation, range-based, and volume-based metrics ((C.1), (C.2), (C.3), (C.4)) provide a comprehensive evaluation by accounting for temporal structure and anomaly duration (Hundman et al., 2018; Blazquez-Garcia et al., 2019).

Appendix D. Reconstruction loss

See Fig. D.2.

Appendix E. Reconstruction loss vs window size

See Fig. E.3.

E.1. Sparse attention impact for short and long sequences

See Fig. E.4 and Table E.9.

Appendix F. Sparse attention computational redundancy elimination vs performance penalty

Fig. F.5 contains two different scenarios that illustrate distinct anomaly patterns commonly encountered in time series data.

The **spike anomaly scenario** depicts a sudden, short-duration deviation where signal values abruptly surge beyond normal operational parameters before rapidly returning to baseline behavior. Figs. F.5(a) and F.5(b) demonstrate how Sparse Attention maintains exceptional detection performance despite significantly reduced computational demands compared to Full Attention. This selective attention approach reduces computational complexity by approximately 90% (from 5.12 MFLOPs to 0.52 MFLOPs) while achieving more precise anomaly detection. The visualization reveals that Sparse Attention (solid green line) produces sharper, more responsive anomaly scores precisely aligned with the ground truth anomaly at time steps 65–75 (highlighted in red), while Full Attention (dashed blue line) generates a more erratic response with higher fluctuations throughout the time series. In the anomaly scores comparison, both approaches detect the true anomaly, with Sparse Attention showing a cleaner signal that more clearly rises above its threshold (dotted green line) during the anomaly period and remains below threshold during normal operation. In contrast, Full Attention shows more variability throughout the time series with its scores frequently approaching its threshold (dotted blue line) even outside the anomaly period.

In contrast, the **trend anomaly** scenario represents a gradual, sustained directional shift where values progressively deviate from expected patterns over an extended period, indicating systematic drift rather than instantaneous disruption. As shown in Figs. F.5(c) and F.5(d), Sparse Attention achieves remarkable efficiency while using fewer computational resources. The time series visualization shows a clear upward trend in the signal (blue line) that begins around time step 65 and persists until approximately time step 115, as highlighted by the red shaded area marking the ground truth anomaly. During this period, the signal shifts from its baseline of approximately −0.5 to a sustained level around 0.3, with an initial spike reaching nearly 1.0 at the onset of the anomaly.

The anomaly scores comparison illustrates that both Sparse Attention (solid green line) and Full Attention (dashed blue line) successfully detect the trend anomaly, with their scores rising above their respective thresholds (dotted green and blue lines) throughout the anomaly period. Notably, both approaches show very similar performance during the actual anomaly, with scores consistently near 1.0 while the trend persists. However, after the anomaly ends (around time step 115), Full Attention shows a lingering elevated score around time step 125, indicating a slight delay in recognizing the return to normal conditions, while Sparse Attention returns more quickly to below-threshold values. The computational complexity comparison illustrates how the sparse mechanism reduces computational requirements from 5.12 MFLOPs to 0.67 MFLOPs (an 87% reduction) while maintaining comparable detection performance. This significant efficiency gain is achieved through the Sparse Attention model's ability to establish selective connections between current observations and key historical reference points, creating an adaptive temporal baseline against which gradual shifts become readily apparent.

The results obtained from these two different scenarios demonstrate how Sparse Attention overcomes Full Attention due to its selective computational approach that delivers both significantly reduced resource requirements and improved detection precision. The Sparse Attention mechanism's ability to focus computation exclusively on relevant patterns makes it particularly effective for various sequence lengths.

Appendix G. Implementation details

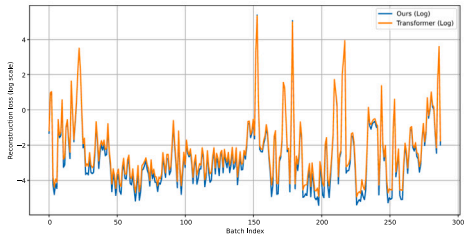
The MAAT model was implemented using **Python 3.8** and the **PyTorch** deep learning framework. PyTorch was chosen for its flexibility in defining custom architectures, efficient tensor operations, and strong support for GPU acceleration. The model was trained and evaluated on an **NVIDIA RTX Titan GPU with 24 GB VRAM**, enabling efficient parallel computation for high-dimensional time-series data.

To ensure stable and reproducible training, experiments were conducted in a controlled environment with fixed random seeds. The training process leveraged mixed-precision computation to optimize memory usage and speed, particularly beneficial for large-scale datasets such as SMD and SWaT. The implementation also utilized PyTorch's `DataLoader` for efficient batch processing, automatic differentiation via `autograd`, and hardware acceleration with `torch.cuda`.

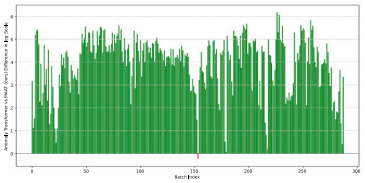
The model was optimized using the **Adam optimizer** with weight decay regularization, and learning rate scheduling was applied to adapt to convergence dynamics. All experiments were run on a system equipped with **Intel Xeon CPUs** and **64 GB RAM** to handle large dataset preprocessing and training workloads efficiently.

Appendix H. MAAT algorithm

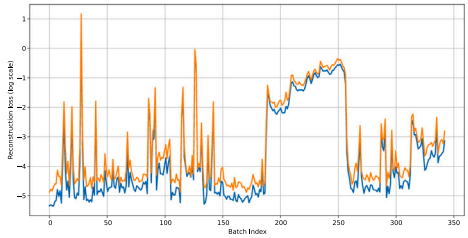
This appendix contains details about the MAAT block algorithm used in the study.



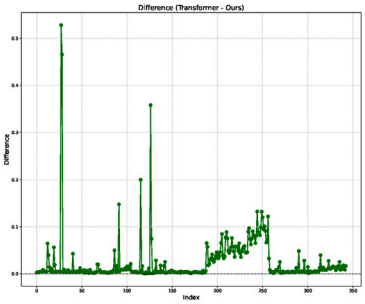
(a) MSL Reconstruction Loss (Ours vs. Anomaly Transformer)



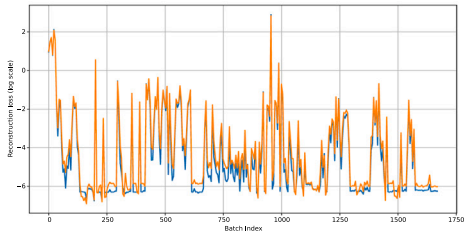
(b) MSL Difference in Reconstruction Loss



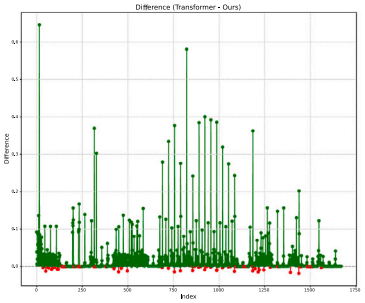
(c) PSM Reconstruction Loss (Ours vs. Anomaly Transformer)



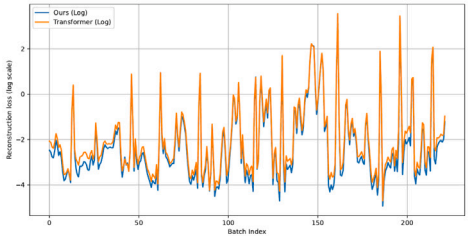
(d) PSM Difference in Reconstruction Loss



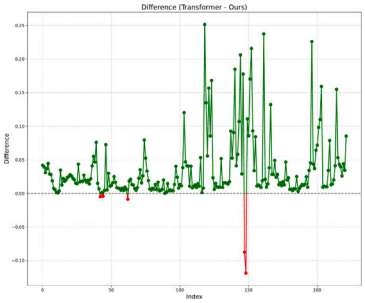
(e) SMAP Reconstruction Loss (Ours vs. Anomaly Transformer)



(f) SMAP Difference in Reconstruction Loss

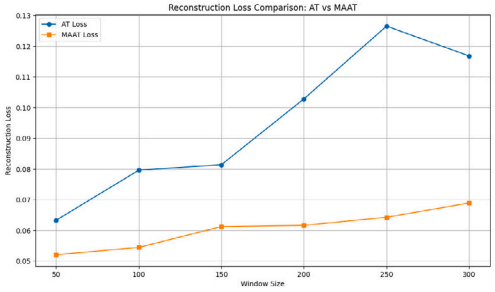


(g) SMD Reconstruction Loss (Ours vs Anomaly Transformer)

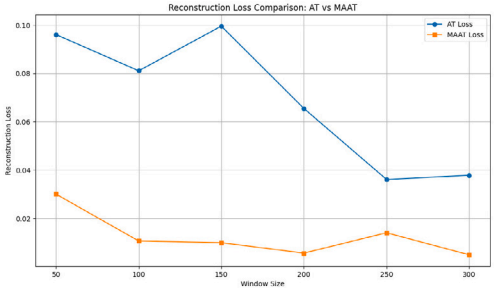


(h) SMD Difference in Reconstruction Loss

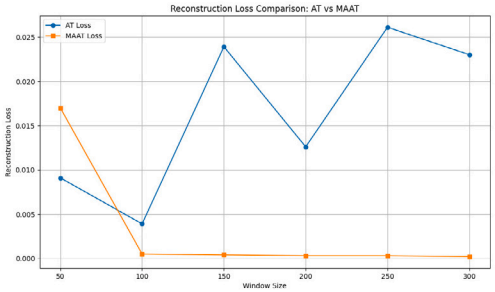
Fig. D.2. Comparison of reconstruction loss between our model and the Anomaly Transformer across different datasets. Left column: Reconstruction loss curves for both models. Right column: Difference in reconstruction loss.



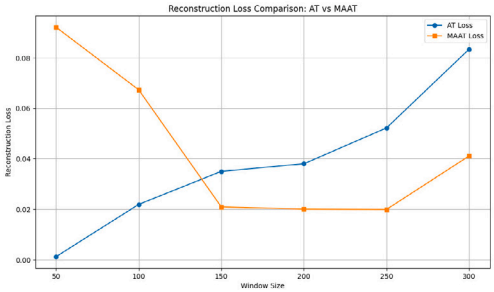
(a) SMD Reconstruction Loss (Ours vs. Anomaly Transformer)



(b) MSL Reconstruction Loss (Ours vs. Anomaly Transformer)



(c) SMAP Reconstruction Loss (Ours vs. Anomaly Transformer)



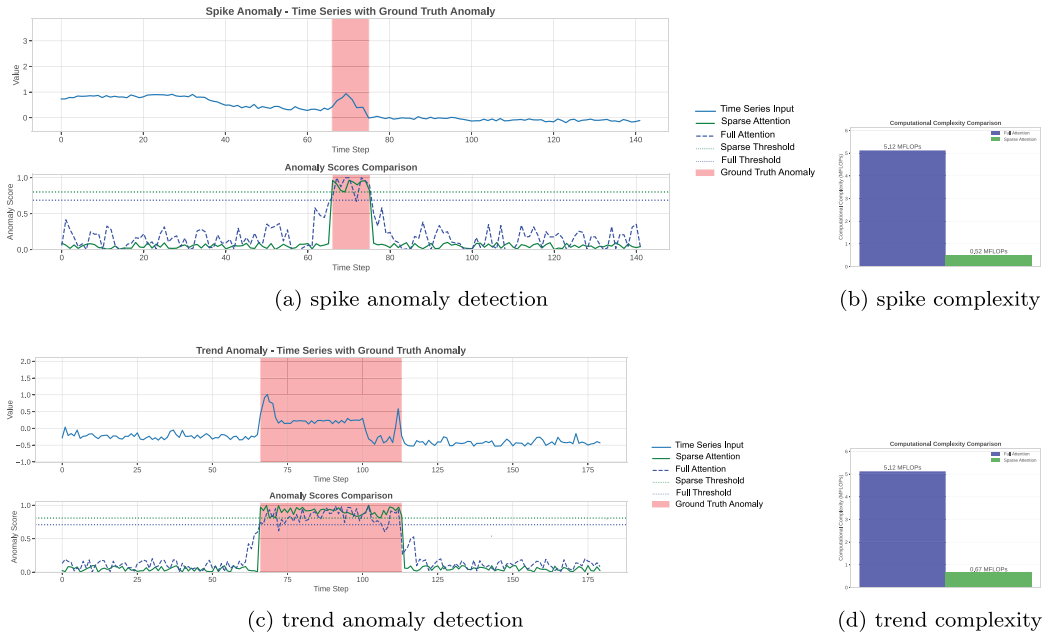
(d) SWaT Reconstruction Loss (Ours vs. Anomaly Transformer)

Fig. E.3. Reconstruction loss sensitivity across different window sizes comparison between our model and the Anomaly Transformer across different datasets.

(a) SMD Reconstruction Loss (Anomaly Transformer vs Anomaly Trasnformer + Sparse Attention)

(b) MSL Reconstruction Loss (Anomaly Transformer vs Anomaly Trasnformer + Sparse Attention)

Fig. E.4. Reconstruction loss comparison on two different datasets; AT and AT+SA refer to Anomaly Transformer and Anomaly Transformer + Sparse Attention respectively.



(a) spike anomaly detection (b) spike complexity

(c) trend anomaly detection (d) trend complexity

Fig. F.5. Sparse attention Computational Redundancy Elimination vs Performance Penalty.

Algorithm 1 MAAT: Mamba Adaptive Anomaly Transformer with Sparse Attention.

Require: Time series $x \in \mathbb{R}^{B \times L \times D}$
Require: Parameters: *block_size*, *d_model*, *n_heads*, *e_layers*, *d_state*, *d_conv*
Ensure: Reconstruction x, Anomaly scores: *series*, *prior*, σ

Initialize $series\_list \leftarrow \emptyset$, $prior\_list \leftarrow \emptyset$, $\sigma\_list \leftarrow \emptyset$
$x_{orig} \leftarrow x$ ▷ Preserve initial input
for $i = 1$ to e_layers **do**
  Sparse Attention Processing:
  $x, series, prior, \sigma \leftarrow \text{sparse\_attn\_layer}(x, block\_size, attn\_mask)$ ▷ Compute sparse block-wise attention over windows of size *block_size*
  Mamba Skip Path:
  $x_{mamba} \leftarrow \text{MambaBlock}(x)$ ▷ State-space model capturing long-range dependencies
  $x_{skip} \leftarrow x_{mamba} + x_{orig}$ ▷ Residual connection
  $x_{skip} \leftarrow \text{LayerNorm}(x_{skip})$
  Adaptive Gating:
  $g \leftarrow \sigma\big(\text{Linear}([x; x_{skip}])\big)$ ▷ Gated feature fusion
  $x \leftarrow g \odot x_{skip} + (1 - g) \odot x$ ▷ Learnable blend of skip and main path
  State Update:
  $x_{orig} \leftarrow x$ ▷ Update input for next layer
  Append *series* to *series_list*
  Append *prior* to *prior_list*
  Append σ to σ_list
end for
if norm $\neq$ None **then**
  $x \leftarrow \text{LayerNorm}(x)$
end if
return $x, series\_list, prior\_list, \sigma\_list$

Data availability

Data used are publicly available, and code repository is indicated in cover letter.

References

Benali, A.A.E., Cafaro, M., Epicoco, I., Pulimeno, M., Schioppa, E.J., 2024. Just in time transformers. IEEE Access 12, 178751–178767. http://dx.doi.org/10.1109/ACCESS.2024.3504862.

Benschoten, A.V., Ouyang, A., Bischoff, F., Marrs, T., 2020. MPA: a novel cross-language API for time series analysis. J. Open Source Softw. 5 (49), 2179. http://dx.doi.org/10.21105/joss.02179.

Blazquez-Garcia, A., Ruiz, J., Pazos, I., Lozano, J.A., 2019. Multivariate anomaly detection in time series data using causal convolutional networks. IEEE Access 7, 130463–130473.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput. Surv. 41 (3), 15:1–15:58.

Child, R., Gray, S., Radford, A., Sutskever, I., 2019. Generating long sequences with sparse transformers. URL https://arxiv.org/abs/1904.10509.

Distante, C., Fineo, L., Mainetti, L., Manco, L., Taccardi, B., Vergallo, R., 2022. HF-SCA: Hands-free strong customer authentication based on a memory-guided attention mechanisms. J. Risk Financ. Manag. 15 (8), 342.

Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Gu, A., Dao, T., Suresh, A.T., Ré, C., 2022. Efficiently modeling long sequences with structured state spaces. In: Advances in Neural Information Processing Systems (NeurIPS). URL https://arxiv.org/abs/2111.00396.

Guo, Y., Zhang, H., Liu, X., 2024. Efficient sparse attention for long sequence time-series forecasting. Sci. Rep. 14, 1–10. http://dx.doi.org/10.1038/s41598-024-66886-1.

Gupta, A., Dar, G., Goodman, S., Ciprut, D., Berant, J., 2021. Memory-efficient transformers via top-k attention. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 5796–5809.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. In: Science, 313, (5786), American Association for the Advancement of Science, pp. 504–507.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Huang, S., Li, X., Hu, W., Liu, S., Peng, W., He, X., 2021. Practical approach to anomaly detection in multivariate time series with missing values. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, pp. 2162–2170.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Soderstrom, T., 2018. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 387–395.

Jeong, Y., Yang, E., Ryu, J.H., Park, I., Kang, M., 2023. Anomalybert: Self-supervised transformer for time series anomaly detection using data degradation scheme. arXiv preprint arXiv:2305.04468.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kitaev, N., Kaiser, L., Levskaya, A., 2020. Reformer: The efficient transformer. Int. Conf. Learn. Represent. (ICLR) URL https://arxiv.org/abs/2001.04451.

Lai, G., Chang, W., Yang, Y., Liu, H., 2021. Revisiting deep learning for time series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Li, D., Chen, D., Jin, B., Shi, L., Goh, J., Ng, S.-K., 2019. MAD-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In: International Conference on Artificial Neural Networks. Springer, pp. 703–716, URL https://arxiv.org/abs/1901.04997.

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation forest. 2008 Eighth IEEE Int. Conf. Data Min. 413–422.

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., Shroff, G., 2016. LSTM-based encoder-decoder for multi-sensor anomaly detection. In: Proceedings of the 2016 International Conference on Machine Learning and Applications. ICMLA, IEEE, pp. 409–414.

Mathur, A., Tippenhauer, N.O., 2016. SWaT: A water treatment testbed for research and training on ICS security. In: 2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater). IEEE, pp. 31–36.

Miller, D.J., Nagaraj, A., Gerdes, R., Rieger, C., 2018. Anomaly detection in drinking water quality data from a real-world water distribution system. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. ACM, pp. 157–158.

NASA, 0000. Space Weather HMI Active Region Patches (SHARP) dataset, Available at: https://ntrs.nasa.gov/citations/20150003032.

NASA, 2018a. Mars Science Laboratory (MSL) dataset, Available at: https://github.com/nasa/telemanom.

NASA, 2018b. Soil Moisture Active Passive (SMAP) dataset, Available at: https://github.com/nasa/telemanom.

Rabiner, L., Juang, B.-H., 1986. Introduction to hidden Markov models. IEEE ASSP Mag. 3 (1), 4–16.

Radford, B.J., Apolonio, L.M., Trias, A.J., Simpson, J.A., 2018. Network traffic anomaly detection using recurrent neural networks. URL https://arxiv.org/abs/1803.10769.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. F-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44.

Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2000. Support vector method for novelty detection. In: Advances in Neural Information Processing Systems. pp. 582–588.

Somepalli, G., Wu, Y., Balaji, Y., Vinzamuri, B., Feizi, S., 2021. Unsupervised anomaly detection with adversarial mirrored AutoEncoders. In: Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence. pp. 1610–1619, URL https://arxiv.org/abs/2003.10713.

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D., 2019. 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. ACM, pp. 2828–2837.

Taieb, S.B., Bontempi, G., Atiya, A.F., Sorjamaa, A., 2012. Machine learning strategies for time series forecasting. Lect. Notes Bus. Inf. Process. 138, 62–77.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Xu, H., Chen, Y., Zhao, W., Bu, J., Li, C., Chen, D., Yu, W., 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. URL https://arxiv.org/abs/1802.03903.

Xu, J., Wu, H., Wang, J., Long, M., 2022. Anomaly transformer: Time series anomaly detection with association discrepancy. In: International Conference on Learning Representations. URL https://openreview.net/forum?id=LzQQ89U1qm_.

Yang, C., Wang, T., Yan, X., 2023a. DDMT: Denoising diffusion mask transformer models for multivariate time series anomaly detection. arXiv preprint arXiv:2310.08800.

Yang, Y., Zhang, C., Zhou, T., Wen, Q., Sun, L., 2023b. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. arXiv preprint arXiv:2306.10347.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., Ahmed, A., 2020. Big bird: Transformers for longer sequences. Adv. Neural Inf. Process. Syst. (NeurIPS) URL https://arxiv.org/abs/2007.14062.

Zhang, Y., Rangapuram, S.S., Wang, Y., Chen, C., Smola, A., 2021. Multi-task time series forecasting with shared attention. arXiv preprint arXiv:2101.09645.

Zhang, K., Wang, Z., Zhou, J., Wang, C., 2023. Gated attention mechanisms for time series forecasting. IEEE Trans. Neural Networks Learn. Syst. 34 (6), 1234–1245. http://dx.doi.org/10.1109/TNNLS.2023.10637425.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W., 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. 35, (12), pp. 11106–11115.