See it, Think it, Sorted: Multimodal Large Language Models are Few-shot Time Series Anomaly Analyzers

Anonymous ACL submission

Abstract

Time series anomaly detection (TSAD) has become increasing important across diverse domains. In TSAD task, while Large Language Models (LLMs) have demonstrated remarkable generalization, few-shot reasoning capabilities in time series tasks, they still fail to match the performance of task-specific methods due to the inherent numerical insensitivity of LLMs' textual tokenizers. With the advancement of LLMs, Multimodal Large Language Models (MLLMs) have emerged as promising candidates for addressing TSAD. Leveraging their exceptional visual reasoning capabilities, MLLMs might analyze time series data by interpreting it in a visual modality, such as plotted graphs, mimicking the way humans perceive and understand visualized information. In this paper, we introduce TAMA, a novel framework that pioneers the integration of MLLMs' image-modality reasoning capabilities into TSAD. Experimental results demonstrate that TAMA's design significantly ehances MLLMs in TSAD task, achieving state-of-the-art performance. Additionally, we contribute one of the first open-source datasets featuring both anomaly classification labels and contextual descriptions, thereby facilitating broader exploration and advancement in this critical field. Our code[1] and dataset[2] have been anonymously open-sourced.

1 Introduction

Time series data has been rapidly proliferating across diverse domains such as finance (Yu et al., 2023), manufacture (Scholz et al., 2024), and industrial monitoring (Feng et al., 2019). Anomalies (defined as unexpected deviations from typical patterns) can signal critical events such as device malfunctions and system failures(Chen et al., 2024a; Nam et al., 2024), which underscored the critical need for robust time series anomaly detection (TSAD) techniques.

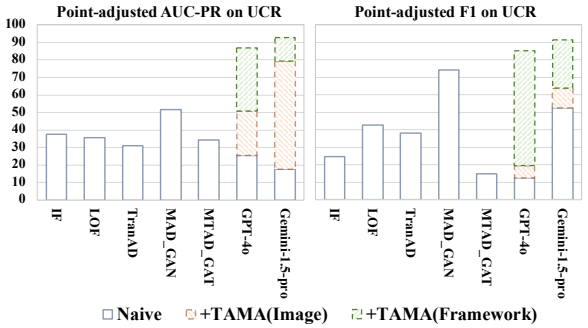


Figure 1: Comparison of AUC-PR and F1 scores with point-adjusted evaluation on the UCR dataset. The models include traditional ML methods (IF, LOF), DL approaches (TranAD, MAD-GAN, MTAD-GAT), and MLLMs (GPT-4o, Gemini-1.5-pro), enhanced with our proposed TAMA framework.

The emergence of Large Language Models (LLMs) has introduced a transformative paradigm shift in TSAD, transitioning the focus from conventional task-specific methods to the exploration and application of LLMs for addressing TSAD challenges. Traditional machine learning (ML) methods often rely on strong assumptions (such as data stationarity), or require handcrafted features (Liu et al., 2008; Feng et al., 2019; Ramaswamy et al., 2000; Yairi et al., 2001; Chen and Guestrin, 2016; Huang et al., 2013). Alternatively, mainstream deep learning (DL) techniques (Liu et al., 2023; Chalapathy and Chawla, 2019; Pang et al., 2022) have limited generalizing ability and explainability (Jacob et al., 2021) due to their reliance on extensive parameter tuning, anomaly-free training data, and black-box detection methods. Consequently, the remarkable task generalization and few-shot learning capabilities of LLMs (Brown et al., 2020a; Gruver et al., 2024; Naveed et al., 2023; Min et al., 2023) have garnered significant attention from re-

[1] https://anonymous.4open.science/r/TAMA-74E1/

[2] https://drive.google.com/drive/folders/1G6a9RxJ-Fwn9pXuZSKdD3KO3yvN7qyGM

searchers, prompting their application to TSAD and other time series tasks by encoding time series data as textual input for LLMs (Jin et al., 2023; Su et al., 2024; Gruver et al., 2024; Liu et al., 2024). However, empirical studies have revealed that without further adaptation, LLMs often fall short of achieving performance levels comparable to those of task-specific methods in time series applications, particularly in the domain of TSAD (Elhafsi et al., 2023; Alnegheimish et al., 2024; Merrill et al., 2024).

The limitations of LLMs in TSAD primarily stem from the insensitivity of their textual tokenizers to numerical values (Qian et al., 2022; Ye et al., 2024), which significantly restricts their ability to detect anomalies involving subtle amplitude changes (Choi et al., 2021). This observation suggests that textual input may not be the optimal format for processing time series data. In response, recent advancements in Multi-modal Large Language Models (MLLMs) (Team et al., 2024; OpenAI et al., 2024), offer promising alternatives. These state-of-the-art models exhibit human-like capabilities in interpreting visual data, including plots and charts, thereby demonstrating advanced proficiency in data analysis tasks (Wang et al., 2024a; Zhang et al., 2024a,b). Notably, just as humans naturally rely on visual graphs rather than raw numerical data to analyze time series, MLLMs can potentially leverage their vision encoders to better interpret plotted time series.

Building on these insights, we propose the central research question: *Can MLLMs be effectively applied to TSAD? If so, what is the appropriate approach to utilize them for this purpose?*

To address this, we present the Time-series Anomaly Multimodal Analyzer (TAMA), a novel framework designed to bridge the gap between TSAD and the capabilities of MLLMs. TAMA leverages MLLMs' multimodal strengths by transforming time series into visual representations ("*see it*"), leveraging a unique three-stage multimodal reasoning mechanism ("*think it*"), and providing accurate anomaly classification alongside contextual explanations and preliminary root cause analysis ("*sorted*"). This structured approach enables TAMA to effectively harness MLLMs' capabilities for robust, interpretable, and generalizable anomaly detection.

Our main contributions are threefold. (1) A novel MLLM-based framework for TSAD: TAMA effectively harnesses the full potential of MLLMs for TSAD, overcoming the direct application limitations of LLMs and existing TSAD methods. (2) An open-sourced dataset: we have constructed one of the first multimodal datasets, providing anomaly detection labels, classification labels, and contextual descriptions associated with time series data, enabling systematic evaluation of our approach. (3) Performance and interpretability improvements: as preliminarily illustrated in Figure 1, extensive experiments show that TAMA outperforms state-of-the-art methods across various TSAD datasets. Furthermore, it enables anomaly classification, detailed descriptions, and preliminary root cause analysis, which collectively highlight its practical utility in TSAD.

2 Related Work

2.1 Time Series Anomaly Detection.

Many surveys (Chalapathy and Chawla, 2019; Pang et al., 2022; Blázquez-García et al., 2021; Choi et al., 2021) are available in the field of TSAD. Classical methods (Ramaswamy et al., 2000; Yairi et al., 2001; Chen and Guestrin, 2016), have established strong baselines in TSAD and continue to be widely used in industry (Wu and Keogh, 2021; Rewicki et al., 2023; Usmani et al., 2022). Deep learning methods (Li et al., 2019; Audibert et al., 2020; Zhang et al., 2018; Su et al., 2019; Zhao et al., 2020; Tuli et al., 2022; Deng and Hooi, 2021; Chen et al., 2024a) focus on learning a comprehensive representation of the entire time series by reconstructing the original input or forecasting through latent variables.

While traditional ML methods require extensive feature engineering (Chalapathy and Chawla, 2019) and DL approaches struggle with generalization and subtle anomalies (Lee et al., 2023), our proposed method addresses these limitations by leveraging image conversion and MLLMs' generalization capabilities.

2.2 LLMs for time series.

Recent work has explored LLMs for time series tasks (Jin et al., 2023; Su et al., 2024; Li et al., 2024; Elhafsi et al., 2023), including innovative approaches like time series tokenization (Gruver et al., 2024), and many other strategies attempting to transfer LLMs' knowledge into time series tasks (Jin et al., 2023; Liu et al., 2024; Alnegheimish et al., 2024). However, these methods are limited by LLMs' inherent constraints with numerical data

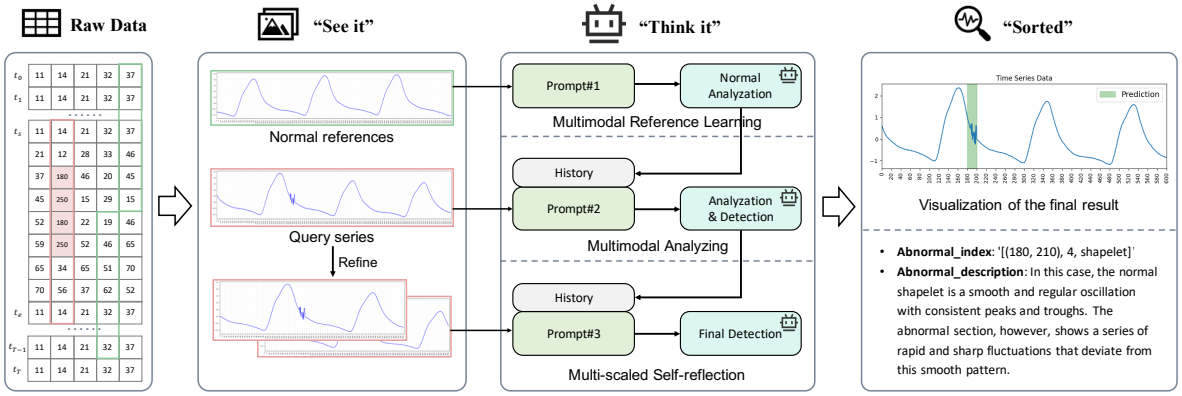


Figure 2: Our framework converts time series into images for visual interpretation ("*See it*"). Then, MLLMs are employed to analyze the visualized time series through Multimodal Reference Learning, Multimodal Analyzing, and Multi-scaled Self-reflection, ensuring self-consistency and stability in the analysis ("*Think it*"). Finally, the detected anomalous intervals are processed into the output format required for TSAD, providing descriptions and possible reasons for each anomaly ("*Sorted*").

(Qian et al., 2022; Ye et al., 2024).

MLLMs show promise in multimodal reasoning (Wang et al., 2024a; Zhang et al., 2024c), and contemporaneous work has discovered that MLLMs can perform TSAD better with input in the visual modality (Wimmer and Rekabsaz, 2023; Daswani et al., 2024; Lin et al., 2024). To the best of out knowledge, we are the first to purpose a systematical MLLM-based framework that surpasses existing TSAD methods. Moreover, our framework innovatively provides comprehensive classification and description to detected anomalies, establishing a new TSAD paradigm. Due to space limitations, please refer to A.1 for a more detailed discussion of related work.

3 Methodology

3.1 Preliminary

Problem Formulation. Consider a time series data $\mathbf{x} = (x_1, x_2, \ldots, x_T) \in R^T$, where x_t represents the sampled value at timestamp t, T refers to the number of timestamps. Throughout the paper, we assume all time-series data to be univariate by default, with multivariate data being transformed into multiple univariate series as needed.

The goal of TSAD is to identify anomalous points or intervals within the time series $\mathbf{x}$. Specifically, an model outputs a sequence of anomaly scores $\mathbf{s} = (s_t)_{t=1}^T = (s_1, s_2, \ldots, s_T) \in R^T$, where s_t indicates the anomaly score corresponding to the data point x_t. By setting a threshold, the anomaly score $\mathbf{s}$ is converted into a set of predicted anomalous intervals $\mathcal{A}_P = \{(t_s, t_e)_{\mathrm{P}}^i\}_{i=1}^{m_P}$, where $(t_s, t_e)^i$ represents the i^{th} out of m_p anomaly intervals, and t_s, t_e indicate the starting and ending indices of each anomaly intervals.

The anomaly classification is a multi-class classification task designed to categorize identified anomalous points into specific types. The output of classification is a set of types $\mathcal{Y}_k = \{y_i\}_{i=1}^{m_k}$, where y_i corresponds to the anomaly classification result for the interval $(t_s, t_e)^i$.

Preprocessing. We preprocess the time series using mean-variance normalization, resulting in normalized data $\overline{\mathbf{x}} = (\overline{x}_1, \ldots, \overline{x}_T)$, where $\overline{x}_t = \frac{x_t - \mu(\mathbf{x})}{\sigma(\mathbf{x})}$, with $\mu(\cdot)$ and $\sigma(\cdot)$ denoting the mean and standard deviation, respectively. Then we utilize overlapped sliding windows to segment the normalized data into a collection of sequence segmentations $\mathcal{P} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_l\}$, where $\hat{\mathbf{x}}_k = (\overline{x}_{k \cdot L_s}, \ldots, \overline{x}_{k \cdot L_s + L_w - 1})$. L_s and L_w are hyperparameters representing the step size and width of the sliding window, respectively. We set overlap ratio $r_o = L_s / L_w < 1$ to allow the same segment of the sequence to be considered across multiple windows. To input time-series data into MLLMs, we plot each sliding window using Matplotlib. Further details can be found in Appendix A.3.2.

Post-processing. Point adjustment (PA) is a widely used post-processing method in TSAD tasks (Kim et al., 2021; Blázquez-García et al., 2021). The PA works as follows: if at least one moment in a contiguous anomaly segment is detected as an anomaly, the entire segment is then considered to be correctly predicted as anomaly. However, this adjustment tends to overestimate model performance. Therefore, to evaluate models more ac-

curately, we adopt a threshold-based PA method instead. The PA with threshold α is defined as:

$$\mathcal{A}_{\text{PA}}(\alpha) = \mathcal{A}_{\text{P}} \cup \{t | t \in (t_s, t_e)^i_{\text{T}}, \\ |(t_s, t_e)^i_{\text{T}} \cap (t_s, t_e)^j_{\text{P}}| > \alpha \cdot L((t_s, t_e)^i_{\text{T}})\}, \quad (1)$$

where α refers to the point-adjustment threshold (PAT) from 0 to 1, where 0 equals PA $\mathcal{A}_{\text{PA}}$ and 1 represents original prediction $\mathcal{A}_{\text{P}}$, and the $L((t_s, t_e)^i_{\text{T}})$ refers to the length of $(t_s, t_e)^i_{\text{T}}$.

3.2 Time Series Anomaly Analyzer (TAMA)

The proposed TAMA framework is illustrated in Figure 2. TAMA comprises three sections: *Multimodal Reference Learning*, *Multimodal Analyzing*, and *Multi-scaled Self-reflection*. The prompts are all available in Appendix A.3.4.

Multimodal Reference Learning. The wide variety of anomaly patterns in time series datasets poses a significant challenge for model adaptation. To address this challenge, we leverage the in-context learning (ICL) capability of MLLMs, which allows them to analyze contextual information and adapt more effectively to different data distributions (Brown et al., 2020b). Specifically, we provide the MLLMs with a set of normal images $\mathcal{I} = \{\mathbf{I}_i \mid i \in \{1, \ldots, n_r\}\}$, where n_r denotes the number of reference images. These reference images represent normal sequences without anomalies, and will be used in subsequent sections, enhancing the analysis of TAMA.

Multimodal Analyzing. In practical TSAD scenarios, providing more interpretable information alongside the detection results is crucial to effectively guide the process of addressing detected anomalies. To this end, we propose a multimodal analysis mechanism within our framework, which complements each anomaly detection result with analyzing, including classification and descriptive contextual information. For the k^{th} window, the MLLM processes image input with corresponding prompts to generate a set of anomaly intervals $\mathcal{A}_k = \{(t_s, t_e)^i\}_{i=1}^{m_k}$ and their classification $\mathcal{Y}_k = \{y_i\}_{i=1}^{m_k}$, where y_i, where y_i denotes the classification for interval $(t_s, t_e)^i$. Guided by prompts, the MLLM also generates anomaly descriptions $\mathcal{T}_k = \{\text{E}_i\}_{i=1}^{m_k}$ and confidence scores $\mathcal{C}_k = \{c_i\}_{i=1}^{m_k}$. As illustrated in Figure 2, the complete model output across all N sliding windows is represented as: $\mathcal{Z}_{\text{raw}} = \{(\mathcal{A}_k, \mathcal{Y}_k, \mathcal{C}_k, \mathcal{T}_k)\}_{k=1}^{N}$, where N is the total number of sliding windows.

Multi-scaled Self-reflection. Due to the constraints on the maximum image size that MLLMs can process, plotted images of longer time series intervals inherently contain richer semantic information and can significantly make entire system more efficient. However, this comes at the cost of temporal compression, making it more challenging to preserve the precise shape of the signal. At a coarser temporal scale, a critical issue arises where periodic peaks may appear abnormally sharp, potentially resembling point anomalies and leading to frequent false positive detections.

To address this challenge, we designed TAMA to incorporate a self-reflection mechanism aimed at improving detection accuracy. This mechanism allows MLLMs to re-evaluate anomalous intervals by viewing them at a higher temporal resolution, enabling better differentiation between true anomalies and artifacts caused by temporal compression. When an anomaly is identified within a segment $\hat{\mathbf{x}}_i$, the segment is further processed using a sliding window approach to generate K overlapping finer-grained segments, denoted as $\{\hat{\mathbf{x}}_i^k\}_{k=1}^{K}$. Each finer segment, $\hat{\mathbf{x}}_i^k$, represents a locally zoomed-in version of the original input segment $\hat{\mathbf{x}}_i$, providing a more detailed representation of the data. This mechanism enables TAMA to reassess its detection decision at a finer scale, allowing for more precise and reliable anomaly identification.

Results Aggregation. The aggregation process begins by mapping local interval indices to a global index space and then aggregates predicted confidence scores across all intervals, where points appearing in multiple intervals receive a summed confidence score. This aggregation produces a confidence sequence $\tilde{\mathbf{c}} = (\tilde{c}_1, \ldots, \tilde{c}_T)$ that matches the original sequence length. Similarly, a point-wise anomaly classification sequence $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_T)$ is constructed through majority voting across overlapping intervals. The final anomaly set $\mathcal{R} = \{i \mid 1 \leq i \leq T, \tilde{c}_i \geq c_0\}$ is determined by applying a confidence threshold c_0.

4 Experiments

In this section, we seek to answer the question: *Can MLLMs be effectively apllied to TSAD?* by conducting experiments, including Anomaly Detection and Classification. Additionally, we also provide a case study to present the capacity of TAMA.

Table 1: Quantitative results across six datasets use metrics point-adjusted *F1%*, *AUC-PR%*, and *AUC-ROC%*. Best and second-best results are in bold and underlined, respectively.

Dataset	UCR						NASA-SMAP						NASA-MSL					
Metric	F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%	
IF	24.7	77.3	37.7	44.6	24.4	25.3	54.2	94.2	58.9	77.1	65.0	87.7	47.6	88.6	53.6	80.4	68.7	88.7
LOF	42.8	**100**	35.6	50.0	92.8	**99.9**	62.2	**100**	43.4	61.4	60.1	99.9	36.4	66.8	44.5	66.0	58.6	99.8
GDN	71.4	80.6	33.4	59.0	87.1	**99.9**	76.4	**100**	40.8	66.2	86.1	**100**	85.1	**100**	38.7	56.7	93.8	**100**
TranAD	38.2	93.7	30.9	51.0	77.0	**99.9**	59.0	99.6	36.8	73.9	74.4	**100**	64.6	99.1	49.2	79.6	82.5	99.9
AnomalyTransformer	54.2	54.2	42.8	42.8	97.3	97.3	90.0	90.0	93.5	93.5	99.2	99.2	83.6	83.6	93.3	93.3	99.2	99.2
TimesNet	32.8	45.8	15.4	23.5	98.4	99.4	**97.7**	**100**	51.4	90.3	**99.8**	**100**	97.4	**100**	52.9	79.7	**99.8**	**100**
SIGLLM (GPT-4o)	23.1	44.6	7.40	15.5	93.5	96.5	69.0	97.8	29.1	49.2	95.5	99.8	70.7	97.9	72.4	**100**	90.0	**100**
GPT4TS	58.0	58.0	56.4	56.4	88.2	88.2	92.1	92.1	82.1	82.1	94.1	94.1	92.6	92.6	88.6	88.6	96.2	96.2
TAMA	**92.5**	97.6	**93.0**	**97.7**	**99.8**	**99.9**	94.5	**100**	**95.5**	**100**	98.4	**100**	**97.5**	**100**	**99.4**	**100**	**99.8**	**100**

Dataset	SMD						ECG						Dodgers					
Metric	F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%	
IF	**83.9**	**100**	73.8	97.0	99.5	**100**	80.8	**99.0**	73.4	92.2	97.2	**100**	48.4	48.4	52.2	52.2	**89.4**	**89.4**
LOF	27.8	75.2	39.9	64.6	52.9	59.3	21.8	39.8	41.4	60.7	56.3	84.2	45.3	45.3	40.8	40.8	63.0	63.0
GDN	76.9	99.7	55.0	88.6	77.0	**100**	75.2	96.2	76.6	97.4	96.9	99.9	37.0	37.0	31.3	31.3	74.2	74.2
TranAD	77.0	99.6	70.9	91.0	96.8	**100**	69.1	98.9	74.7	**97.7**	94.9	**100**	38.2	38.2	33.9	33.9	74.6	74.6
AnomalyTransformer	32.8	32.8	61.5	61.5	64.7	64.7	31.4	31.4	26.7	26.7	76.8	76.8	37.2	37.2	64.2	64.2	83.8	83.8
TimesNet	82.8	**100**	57.3	99.9	95.4	**100**	**92.4**	96.6	**90.0**	97.6	**99.4**	**100**	48.1	48.1	73.0	73.0	83.7	83.7
SIGLLM (GPT-4o)	42.9	59.8	30.4	53.1	68.8	77.8	19.2	50.4	71.0	87.6	94.2	96.9	48.1	48.1	60.7	60.7	83.2	83.2
GPT4TS	76.1	76.1	81.3	81.3	83.4	83.4	16.1	16.1	59.4	59.4	53.6	53.6	10.2	10.2	50.4	50.4	51.8	51.8
TAMA	77.8	**100**	**87.9**	**100**	98.9	**100**	81.3	87.5	84.5	90.0	95.4	99.4	**65.6**	**65.6**	**74.0**	**74.0**	85.2	85.2

Experimental Settings. We select GPT-4o (OpenAI et al., 2024) as TAMA's default model, and the specific version we used is "gpt-4o-2024-05-13". To ensure the stability of TAMA and the reproducibility of the results, the *temperature* is set to 0.1 and the *top_p* is set to 0.3. Besides, the JSON mode of GPT-4o is used to facilitate subsequent result analysis. The detailed settings and the usage of tokens are presented in Appendix A.3.

Datasets. We use a diverse set of real-world datasets across multiple domains for both anomaly detection and anomaly classification tasks. These domains include web service: SMD (Su et al., 2019), industry: UCR (Wu and Keogh, 2021) and NormA (Boniol et al., 2021), scientific measurement: NASA-SMAP (Hundman et al., 2018) and NASA-MSL (Hundman et al., 2018), health care: ECG (Paparrizos et al., 2022), and transportation: Dodgers (Hutchins, 2006). All datasets are univariate except for SMD. The detailed statistical information of the dataset can be found in Table 9 in the Appendix. We convert SMD into an univariate dataset by splitting it channel-wise for our anomaly detection experiment.

Due to the limited availability of datasets with anomaly classification labels, we created an anomaly classification dataset by combining four real-world datasets (UCR, NASA-SMAP, NASA-MSL, and NormA) with manually labeled anomaly types, along with a synthetic dataset generated using GutenTAG (Wenig et al., 2022). The datasets cover diverse application scenarios, including healthcare, scientific measurements, and industrial records, providing a challenging benchmark for evaluating model generalization. Anomaly type annotations are refined by domain experts. Additionally, the synthetic dataset, derived from Lai et al. (Lai et al., 2021), consists of 7200 samples with three anomaly types: *point*, *trend*, and *frequency*. The visualization of each typical anomaly type is included in Appendix 7.

4.1 Anomaly Detection

Baselines. The baseline models used in our experiments include both MLs (IF (Liu et al., 2008), LOF (Huang et al., 2013)) and DL (AnomalyTransformer (Xu et al., 2021), TranAD (Tuli et al., 2022), GDN (Deng and Hooi, 2021), TimesNet (Wu et al., 2023)) methods. For more DL baselines including MAD_GAN (Li et al., 2019), MSCRED (Zhang et al., 2018), MTAD_GAT (Zhao et al., 2020), OminiAnomaly (Su et al., 2019), and USAD (Audibert et al., 2020), please refer to Table 10 in Appendix A.4.1. Besides, the GPT4TS (Zhou et al., 2023) and SIGLLM (Alnegheimish et al., 2024) are two LLM-based baselines, where SIGLLM is reproduced with GPT-4o. All baseline models has been run with the default configurations. For those datasets without default configurations, we managed to optimize the performance by searching the best parameters.

Metrics. Following the mainstream of TSAD, we evaluated TAMA and other baselines by the point-adjusted *F1*, *AUC-PR*, *AUC-ROC*. Additionally, as

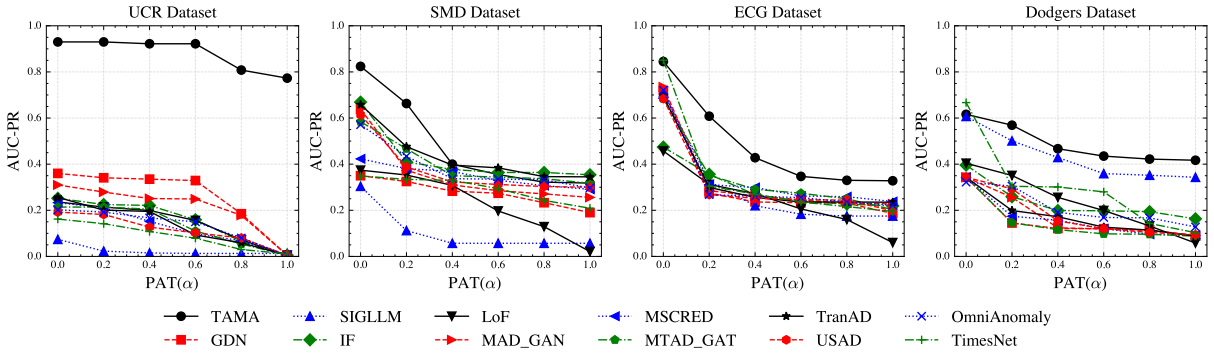


Figure 3: The *AUC-PR* of all models at various point-adjustment threshold α (PAT, defined in Section 3.1).

mentioned in Section 3.1, using PA greatly overestimates the models' performance. To address this, we re-evaluate all models using PA with a threshold α (PAT, defined in Section 3.1).

Main Results. The experimental results are presented in Table 1. The *mean* represents the average performance across all sub-series, while the *maxima* reflects the best performance achieved among all sub-series. In terms of the *maxima* value, our proposed method, TAMA, achieves results comparable to or even exceeding those of baseline models on certain datasets. More importantly, TAMA consistently outperforms nearly all baselines in the *mean* metric, demonstrating a particularly strong advantage on industry and transportation datasets.

To assess the impact of PA, we evaluate model's performance under varying PATs using the *AUC-PR* metric. As shown in Figure 3 (full results in Appendix A.4.2), performance declines for all models as PAT increases, indicating that full PA (α=0) has overestimated the model performance. Nevertheless, TAMA consistently outperforms baseline methods across all PAT settings, demonstrating TAMA's strong robustness and stability.

In summary, results demonstrates that MLLMs can be effectively applied to TSAD tasks through TAMA, our proposed framework. Moreover, these findings confirm that TAMA not only achieves competitive performance but also delivers reliable and stable results, making it a more dependable and promising solution for TSAD.

4.2 Anomaly Classification

In practical applications, it is preferable not only to detect anomaly intervals but also to provide a classification indicating their causes. Thus, we conduct the anomaly classification task.

The overall results presented in Table 2 indicate that TAMA, guided by the provided prompts, demonstrates a reliable understanding of each type of anomaly and can accurately classify most anomalies, with the exception of seasonal anomalies. TAMA performs exceptionally well in classifying shapelet anomalies, suggesting that it effectively captures the shape of the input sequences. However, it is evident that the framework struggles with seasonal anomalies. It may be caused by a lack of relevant materials in the MLLM's pre-training stage, which results in a weak understanding of concepts such as "seasonality" or "frequency".

Table 2: Classification is detailed for each anomaly type, with 'total' representing the overall performance.

Type	Point	Shapelet	Seasonal	Trend	**Total**
Accuracy%	81.0	99.2	29.0	74.5	78.5
Support	100	246	100	94	567

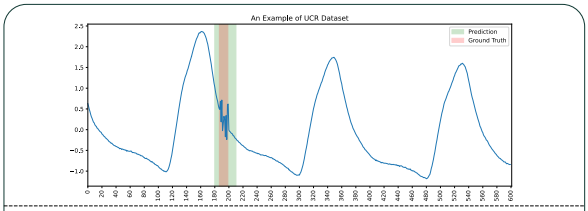
4.3 Case Study

There is a case study in Figure 4 (More case studies are available in Appendix A.4.5). As the figure shown, TAMA can output the anomaly range and its type, while offering an explanation of it. According to the explanation, we can find that TAMA can successfully identify the anomaly by comparing the normal pattern. Moreover, with the background information, TAMA can try to seek the reason of this anomaly.

5 Discussion

As the result shown in Table 1, TAMA outperforms most baselines, demonstrating that MLLMs can be applied to TSAD. In this section, based on TAMA, we seek to answer the question: *What is the appropriate way of applying LLMs or MLLMs to TSAD?*

The visual modality proves more effective than the textual modality for MLLMs in TSAD. As shown in Table 3, we conduct an experiment to compare the visual modality and text modality, while keeping the prompts and procedures as the



Detection :

- **Anomaly range**: (180, 210)
- **Anomaly type**: shapelet

Anomaly Description:

The abnormality observed between indices 180 and 210 is best classified as a 'shapelet' outlier. Shapelet outliers refer to subsequences with dissimilar basic shapelets compared to the normal shapelet. In this case, the normal shapelet is a smooth and regular oscillation with consistent peaks and troughs. The abnormal section, however, shows a series of rapid and sharp fluctuations that deviate from this smooth pattern.

The abnormality in the time series data, characterized by rapid oscillations and fluctuations, could potentially be indicative of an internal bleeding event. In medical monitoring, such irregular patterns might reflect sudden changes in physiological parameters, such as blood pressure or heart rate, which can occur during episodes of internal bleeding. The disruption in the otherwise stable pattern suggests an acute event that requires further investigation to confirm the presence and cause of internal bleeding.

Figure 4: The case study in UCR dataset.

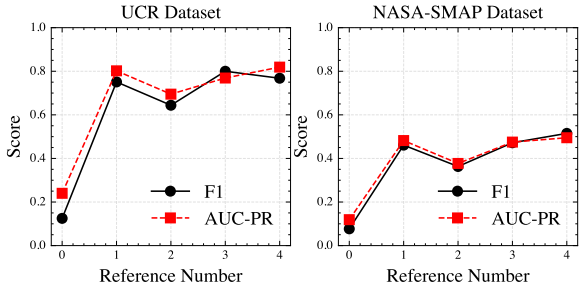


Figure 5: Results (without PA) of reference number ablation experiments.

same. Compared to methods using textual modality, TAMA (Image), has made significant improvement, with a 37.9% increase on NASA-MSL and a 36.9% increase on NASA-SMAP. This result indicates that for anomaly detection tasks, adopting visual modality is more beneficial. The further discussion about this phenomenon is in Appendix A.2.1.

The incorporation of high-quality reference data is crucial for enhancing MLLMs' performance. To investigate the effectiveness of Multimodal Reference Learning, we conduct experiments in two aspects: (1) the number of references. (2) the information in references. Experiments are conducted on UCR and NASA-SMAP datasets. Additionally, to avoid interference from other modules, self-reflection is not added.

In the first experiment, as shown in the Figure 5, results reveal that introducing references can significantly improve TAMA's performance. However, when the number of reference increases, the performance seems to have not obviously improvement.

In the second experiment, we replace the normal data with abnormal data. The *normal* refers to using the normal data as the reference data, while *abnormal* indicates using abnormal data. As the results presented in Table 4, we can find that *normal* performs better than *abnormal* on both UCR and NASA-SMAP datasets, showing that the information of references can notably impact the model's performance, which suggests the MLLM can truly learn normal patterns from the reference data.

In summary, the two experiments show that it is significant to offer some valuable references for MLLMs. With the valuable references, MLLMs can outperforms most tradditional methods.

Multi-scaled Self-reflection can enhance the performance and stability. Experiments are conducted on SMD, NASA-SMAP and NASA-MSL to investigate the impact of the self-reflection. TAMA* represents TAMA without self-reflection. The results shown in Table 5 demonstrate that self-reflection enhances the performance of TAMA. As the complexity of the data increases, the performance improvement becomes more pronounced, validating the effect of self-reflection.

The TAMA framework demonstrates universal effectiveness in improving MLLMs' anomaly detection performance, regardless of the selection of MLLMs. The experiment is conducted on the UCR dataset. For each MLLM, we conduct experiments both with (*+TAMA*) and without (*Naive*) TAMA framework. The experimental results are presented in the Table 6. We use the original *AUC-PR* without point-adjustment as the metric. The findings reveal that all MLLMs exhibit a substantial enhancement in performance on the UCR dataset following their integration into TAMA. This not

Table 3: Performance comparison of image and text modalities, average PA *F1%* as the metric. TAMA (Image) and TAMA (Text) are based on TAMA but using image-modality and text-modality respectively.

Modality	NASA-MSL	NASA-SMAP
TAMA (Image)	97.5	94.5
TAMA (Text)	70.7	69.0
SIGLLM (Text)	42.9	43.1

Table 4: The performance comparison in different reference images. Metric is *AUC-PR%* without PA.

Dataset	*normal*	*abnormal*
UCR	83.0	46.8 (**-36.2**)
NASA-SMAP	72.9	48.5 (**-24.4**)

Table 5: The performance of TAMA and TAMA*, using the *AUC-PR%* without PA as the metric.

Dataset	SMD	NASA-SMAP	NASA-MSL
TAMA	87.9	95.5	99.4
TAMA*	78.6 (**-9.3**)	89.2 (**-6.3**)	97.7 (**-1.7**)

Table 6: Comparison of different pre-trained LMMs using the average *AUC-PR%* without PA as the metric. We compare results with (+***TAMA***) and without our framework (*Naive*).

LMM	*Naive*	+***TAMA***
GPT-4o	41.8	80.2 (**+38.4**)
GPT-4o-mini	11.8	51.1 (**+39.3**)
Gemini-1.5-pro	25.4	87.8 (**+62.4**)
Gemini-1.5-flash	17.9	36.4 (**+18.5**)
qwen-vl-max-0809	61.7	80.5 (**+18.8**)

only validates that TAMA improves the MLLMs' abilities in anomaly detection but also confirms the generalizability of TAMA's framework.

6 Scalability Study

While TAMA demonstrates superior performance compared to most baselines, its MLLM-based approach shows the limitation in scalability compared to some specialized methods. To address this problem, we introduce Principal Component Analysis (PCA) and down-sampling to reduce the scalability.

High-dimensional Data. In TAMA, an input image only contains an univariate time series, which leads to multiple images when dealing with multivariate time series. To address this scalability problem, we employ PCA to reduce the dimensionality of the SMD dataset from 38 to 2 dimensions (the visualization is in Appendix A.4.7). The experimental results presented in Table 7 demonstrate that although there is a slight performance degradation on the reduced SMD dataset, the model preserves 90% of its original performance while achieving a 95% reduction in computation.

High sampling rate Data. High-frequency sampled data often contains redundant information, resulting in excessive length and low information density. Down-sampling is a widely used technique to deal with this issue. We evaluate TAMA on both down-sampled ECG and NormA datasets, where the input is down-sampled and the output is up-sampled by interpolation, preventing the altering of ground-truth. As shown in Table 8,

Table 7: Comparison of TAMA's performance on SMD before and after (denoted as SMD-R) PCA

Dataset	F1%	AUC-PR%	AUC-ROC%
SMD	93.0	97.2	99.7
SMD-R	85.7	93.1	99.3

Table 8: Results of TAMA over various down-sampling rates on down-sampled ECG and NormA datasets. $F1_{relative} = (F1 - F1_{original})/F1_{original}$, where $F1_{orginal}$ is the F1 score without down-sampling.

Dataset	Rate	F1%	AUC-PR%	AUC-ROC%	$F1_{relative}$ %
ECG	1/2	81.2	93.1	97.8	-0.1
	1/3	79.9	84.3	91.3	-1.7
	1/5	69.9	65.6	82.5	-14.0
NormA	1/2	78.5	83.0	97.9	-2.7
	1/3	78.3	75.5	91.8	-2.9
	1/5	65.4	70.9	93.6	-18.9

TAMA's performance on both datasets decreases as the down-sampling rate increases, indicating that down-sampling disrupts the original pattern of the data. Quantitatively, a down-sampling rate at $\frac{1}{2}$ or $\frac{1}{3}$ brings a drop less than 3% in terms of F1 and AUC-ROC, while reducing the scale by 66.6%. Down-sampling can effectively reduce the overall data length with manageable performance trade-off.

7 Conclusion

In this paper, we propose a novel framework named TAMA and conduct comprehensive experiments to prove that the MLLM can be effective applied to TSAD task under TAMA. Our analysis of TAMA's design reveals two key insights: (1) the visual modality is more effetive than texutal modality more effective for MLLMs in TSAD; (2) harnessing MLLMs' few-shot capacity, references and self-reflection can enhance performance and stability. These findings pave the way for applying MLLMs to TSAD or time series analysis.

Limitations

Some limitations should be noted. (1) TAMA primarily relies on pre-trained MLLMs without fine-tuning. (2) While TAMA primarily focuses on univariate time series, this creates scalability challenges. Although we have discussed it in Section 6, significant improvement in scalability is still an important direction for future research. (3) Although TAMA can infer the underlying causes of anomalies, it is restricted to univariate time series analysis.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024. Large language models can be zero-shot anomaly detectors for time series? _eprint: 2405.14755.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. 2020. USAD: UnSupervised Anomaly Detection on Multivariate Time Series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 3395–3404, New York, NY, USA. Association for Computing Machinery. Event-place: Virtual Event, CA, USA TLDR: A fast and stable method called UnSupervised Anomaly Detection for multivariate time series (USAD) based on adversely trained autoencoders capable of learning in an unsupervised way is proposed.

Tharindu R. Bandaragoda, Kai Ming Ting, David Albrecht, Fei Tony Liu, and Jonathan R. Wells. 2014. Efficient Anomaly Detection by Isolation Using Nearest Neighbour Ensemble. In *2014 IEEE International Conference on Data Mining Workshop*, pages 698–705. TLDR: iNNE (isolation using Nearest Neighbour Ensemble), an efficient nearest neighbour-based anomaly detection method by isolation that overcomes three weaknesses of iForest, i.e., Its inability to detect local anomalies, anomalies with a low number of relevant attributes, and anomalies that are surrounded by normal instances.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. 2021. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Comput. Surv.*, 54(3). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. *The VLDB Journal*, 30(6):909–931.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language Models are Few-Shot Learners. _eprint: 2005.14165.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep Learning for Anomaly Detection: A Survey. *arXiv:1901.03407 [cs, stat]*. ArXiv: 1901.03407.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Feiyi Chen, Zhen Qin, Mengchu Zhou, Yingying Zhang, Shuiguang Deng, Lunting Fan, Guansong Pang, and Qingsong Wen. 2024a. LARA: A Light and Anti-overfitting Retraining Approach for Unsupervised Time Series Anomaly Detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 4138–4149, New York, NY, USA. Association for Computing Machinery. Event-place: Singapore, Singapore TLDR: In LARA, the retraining process is designed as a convex problem such that overfitting is prevented and the retraining process can converge fast, and it is mathematically and experimentally proved that when fine-tuning the latent vector and reconstructed data, the linear formations can achieve the least adjusting errors between the ground truths and the fine-tuned ones.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. 2024b. VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters. *arXiv preprint*. ArXiv:2408.17253.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery. Event-place: San Francisco, California, USA.

Kukjin Choi, Jihun Yi, Changhwa Park, and Sungroh Yoon. 2021. Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access*, 9:120043–120065.

Enyan Dai and Jie Chen. 2022. Graph-Augmented Normalizing Flows for Anomaly Detection of Multiple Time Series. *ArXiv*, abs/2202.07857.

Mayank Daswani, Mathias M. J. Bellaiche, Marc Wilson, Desislav Ivanov, Mikhail Papkov, Eva Schnider, Jing Tang, Kay Lamerigts, Gabriela Botea, Michael A. Sanchez, Yojan Patel, Shruthi Prabhakara,

Shravya Shetty, and Umesh Telang. 2024. Plots Unlock Time-Series Understanding in Multimodal Models. _eprint: 2410.02637.

Ailin Deng and Bryan Hooi. 2021. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. In *AAAI Conference on Artificial Intelligence*.

Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa A. D. Nesnas, and Marco Pavone. 2023. Semantic anomaly detection with large language models. *Autonomous Robots*, 47(8):1035–1055.

Muhammad Fahim and Alberto Sillitti. 2019. Anomaly Detection, Analysis and Prediction Techniques in IoT Environment: A Systematic Literature Review. *IEEE Access*, 7:81664–81681. TLDR: This paper presents the results of a systematic literature review about anomaly detection techniques except for these dominant research areas, published from 2000 to 2018 in the application areas of intelligent inhabitant environments, transportation systems, health care systems, smart objects, and industrial systems.

Cheng Feng, Venkata Reddy Palleti, Aditya P. Mathur, and Deeph Chana. 2019. A Systematic Framework to Generate Invariants for Anomaly Detection in Industrial Control Systems. *Proceedings 2019 Network and Distributed System Security Symposium*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2024. Large Language Models Are Zero-Shot Time Series Forecasters. *arXiv preprint*. ArXiv:2310.07820 [cs] TLDR: This work proposes procedures for effectively tokenizing time series data and converting discrete distributions over tokens into highly flexible densities over continuous values and shows how LLMs can naturally handle missing data without imputation through non-numerical text, accommodate textual side information, and answer questions to help explain predictions.

Wei He, Zhiheng Xi, Wanxu Zhao, Xiaoran Fan, Yiwen Ding, Zifei Shan, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Distill Visual Chart Reasoning Ability from LLMs to MLLMs. *arXiv preprint*. ArXiv:2410.18798 [cs] TLDR: This paper proposes Code-as-Intermediary Translation (CIT), a cost-effective, efficient and easily scalable data synthesis method for distilling visual reasoning abilities from LLMs to MLLMs, which serves as an intermediary that translates visual chart representations into textual representations, enabling LLMs to understand cross-modal information.

Tian Huang, Yan Zhu, Qiannan Zhang, Yongxin Zhu, Dongyang Wang, Meikang Qiu, and Lei Liu. 2013. An LOF-Based Adaptive Anomaly Detection Scheme for Cloud Computing. In *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*, pages 206–211. TLDR: This work presents an adaptive anomaly detection scheme for cloud computing based on LOF, which learns behaviors of applications both in training and detecting phase and enables the ability to detect contextual anomalies.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 387–395, New York, NY, USA. Association for Computing Machinery. Event-place: London, United Kingdom TLDR: The effectiveness of Long Short-Term Memory networks, a type of Recurrent Neural Network, in overcoming issues using expert-labeled telemetry anomaly data from the Soil Moisture Active Passive (SMAP) satellite and the Mars Science Laboratory (MSL) rover, Curiosity is demonstrated.

Jon Hutchins. 2006. Dodgers Loop Sensor. Published: UCI Machine Learning Repository.

Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. 2021. Exathlon: a benchmark for explainable anomaly detection over time series. *Proc. VLDB Endow.*, 14(11):2613–2626. Publisher: VLDB Endowment.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2023. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. 2021. Towards a Rigorous Evaluation of Time-series Anomaly Detection. In *AAAI Conference on Artificial Intelligence*.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. 2021. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Daesoo Lee, Sara Malacarne, and Erlend Aune. 2023. Explainable time series anomaly detection using masked latent generative modeling. *Pattern Recognit.*, 156:110826.

Kareth M. León-López, Florian Mouret, Henry Arguello, and Jean-Yves Tourneret. 2022. Anomaly Detection and Classification in Multispectral Time Series Based on Hidden Markov Models. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11.

Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. 2019. MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV*, pages 703–716, Berlin, Heidelberg. Springer-Verlag.

Event-place: Munich, Germany TLDR: The proposed MAD-GAN framework considers the entire variable set concurrently to capture the latent interactions amongst the variables and is effective in reporting anomalies caused by various cyber-intrusions compared in these complex real-world systems.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. UrbanGPT: Spatio-Temporal Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 5351–5362, New York, NY, USA. Association for Computing Machinery. Event-place: Barcelona, Spain.

Chunming Lin, Bowen Du, Leilei Sun, and Linchao Li. 2024. Hierarchical Context Representation and Self-Adaptive Thresholding for Multivariate Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3139–3150.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. TLDR: The use of isolation enables the proposed method, iForest, to exploit sub-sampling to an extent that is not feasible in existing methods, creating an algorithm which has a linear time complexity with a low constant and a low memory requirement.

Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, and B. Aditya Prakash. 2024. LSTPrompt: Large Language Models as Zero-Shot Time Series Forecasters by Long-Short-Term Prompting. *ArXiv*, abs/2402.16132.

Ya Liu, Yingjie Zhou, Kai Yang, and Xin Wang. 2023. Unsupervised deep learning for iot time series. *IEEE Internet of Things Journal*, 10(16):14285–14306.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. *arXiv preprint*. ArXiv:2203.10244 [cs].

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. 2024. Language Models Still Struggle to Zero-shot Reason about Time Series. *ArXiv*, abs/2404.11757.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.*, 56(2). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Youngeun Nam, Susik Yoon, Yooju Shin, Minyoung Bae, Hwanjun Song, Jae-Gil Lee, and Byung Suk Lee. 2024. Breaking the Time-Frequency Granularity Discrepancy in Time-Series Anomaly Detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 4204–4215, New York, NY, USA. Association for Computing Machinery. Event-place: Singapore, Singapore TLDR: A TSAD framework that simultaneously uses both the time and frequency domains while breaking the time-frequency granularity discrepancy is proposed, which outperforms state-of-the-art methods by 12.0–147%, as demonstrated by experimental results.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal S. Mian. 2023. A Comprehensive Overview of Large Language Models. *ArXiv*, abs/2307.06435.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor

Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. _eprint: 2303.08774.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. 2022. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2):1–38. ArXiv:2007.02500 [cs, stat].

John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. *Proc. VLDB Endow.*, 15(8):1697–1711. Publisher: VLDB Endowment TLDR: This work thoroughly studied over one hundred papers to identify, collect, process, and systematically format datasets proposed in the past decades to establish a leaderboard of univariate time-series anomaly detection methods, and concludes that TSB-UAD provides a valuable, reproducible, and frequently updated resource.

Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. 2022. Limitations of Language Models in Arithmetic and Symbolic Induction. _eprint: 2208.05051.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 427–438, New York, NY, USA. Association for Computing Machinery. Event-place: Dallas, Texas, USA.

Ferdinand Rewicki, Joachim Denzler, and Julia Niebling. 2023. Is it worth it? comparing six deep and classical methods for unsupervised anomaly detection in time series. *Applied Sciences*, 13(3).

Jona Scholz, Maike Holtkemper, Alexander Graß, and Christian Beecks. 2024. Anomaly Detection in Manufacturing. In John Soldatos, editor, *Artificial Intelligence in Manufacturing: Enabling Intelligent, Flexible and Cost-Effective Production Through AI*, pages 351–360. Springer Nature Switzerland, Cham.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. *ArXiv*, abs/2402.10350.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski,

Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Sey-

more, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pa-

sumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim,

Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. 2022. A Review of Unsupervised Machine Learning Frameworks for Anomaly Detection in Industrial Applications. In *Intelligent Computing*, pages 158–189, Cham. Springer International Publishing.

Chaoyang Wang and Guangyu Liu. 2024. From anomaly detection to classification with graph attention and transformer for multivariate time series. *Advanced Engineering Informatics*, 60:102357.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024a. Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning. *ArXiv*, abs/2401.06805.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *arXiv preprint*. ArXiv:2406.18521 [cs] TLDR: This work proposes CharXiv, a comprehensive evaluation suite involving 2,323 natural, challenging, and diverse charts from arXiv papers, and reveals a substantial, previously underestimated gap between the reasoning skills of the strongest proprietary model and the strongest open-source model.

Phillip Wenig, Sebastian Schmidl, and Thorsten Papenbrock. 2022. TimeEval: a benchmarking toolkit for time series anomaly detection algorithms. *Proc. VLDB Endow.*, 15(12):3678–3681. Publisher: VLDB Endowment.

Christopher Wimmer and Navid Rekabsaz. 2023. Leveraging Vision-Language Models for Granular Market Change Prediction. _eprint: 2301.10166.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*.

Renjie Wu and Eamonn Keogh. 2021. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*.

Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. 2023. Deep Isolation Forest for Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14. ArXiv:2206.06602 [cs] TLDR: A new representation scheme that utilises casually initialised neural networks to map original data into random representation ensembles, where random axis-parallel cuts are subsequently applied to perform the data partition, encouraging a unique synergy between random representations and random partition-based isolation.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2021. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy.

Takehisa Yairi, Yoshikiyo Kato, and Koichi Hori. 2001. Fault Detection by Mining Association Rules from House-keeping Data.

Chao Ye, Haobo Wang, Liyao Li, Sai Wu, Gang Chen, and Junbo Zhao. 2024. Towards Cross-Table Masked Pretraining for Web Data Mining. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 4449–4459, New York, NY, USA. Association for Computing Machinery. Event-place: Singapore, Singapore TLDR: This work first identifies the crucial challenges behind tabular data pretraining, particularly overcoming the cross-table hurdle, and proposes an innovative, generic, and efficient cross-table pretraining framework, dubbed as CM2, where the core to it comprises a semantic-aware tabular neural network that uniformly encodes heterogeneous tables without much restriction.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, page nwae403. Publisher: Oxford University Press.

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023. Harnessing LLMs for Temporal Data - A Study on Explainable Financial Time Series Forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753, Singapore. Association for Computational Linguistics.

Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and N. Chawla. 2018. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. In *AAAI Conference on Artificial Intelligence*.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024a. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Annual Meeting of the Association for Computational Linguistics*.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. 2024b. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. *Preprint*, arXiv:2407.07053.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. 2024c. Multimodal Self-Instruct: Synthetic Abstract Image and Visual Reasoning Instruction Using Language Model. _eprint: 2407.07053.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. 2020. Multivariate Time-Series Anomaly Detection via Graph Attention Network. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 841–850.

Tian Zhou, Peisong Niu, xue wang, Liang Sun, and Rong Jin. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. In *Advances in Neural Information Processing Systems*, volume 36, pages 43322–43355. Curran Associates, Inc.

A Appendices

Table of Contents

- Appendix A.1: Related work
- Appendix A.2: Detailed analysis
- Appendix A.3: More on experimental setup
- Appendix A.4: Complete experimental results and visualization

A.1 Detailed Related Work

A.1.1 Time Series Anomaly Detection.

Classical methods (Ramaswamy et al., 2000; Yairi et al., 2001; Chen and Guestrin, 2016), especially unsupervised methods such as Isolation Forest (IF) (Liu et al., 2008; Bandaragoda et al., 2014), and Local Outlier Factor (LoF) (Huang et al., 2013) are introduced into TSAD in early stages. There are also variants of these classical ML algorithms like Deep Isolation Forest (DIF) (Xu et al., 2023), which enhances IF by introducing non-linear partitioning. ML methods methods perform exceptionally well on many TASD datasets (Wu and Keogh, 2021; Rewicki et al., 2023), have been applied widely in industry (Usmani et al., 2022), and serve as strong baselines in recent researches.

Among all reconstructing-based models, MAD-GAN (Li et al., 2019) is an LSTM-based network enhanced by adversarial training. Similarly, USAD (Audibert et al., 2020) is an autoencoder-based framework that also utilizes adversarial training. MSCRED (Zhang et al., 2018) is designed to capture complex inter-modal correlations and temporal information within multivariate time series. However, its effectiveness can be constrained by limited training data. OmniAnomaly (Su et al., 2019) addresses multivariate time series by using stochastic recurrent neural networks to model normal patterns, providing robustness against variability in the data. MTAD-GAT (Zhao et al., 2020) employs a graph-attention network based on GRU to model both feature and temporal correlations. TranAD (Tuli et al., 2022), a transformer-based model, utilizes an encoder-decoder architecture that facilitates rapid training and high detection performance. Except reconstructing-based method, GDN (Deng and Hooi, 2021) is a forecasting-based model that utilizes attention-based forecasting and deviation scoring to output anomaly scores. Additionally, LARA (Chen et al., 2024a), is a lightweight approach based on deep variational auto-encoders.The novel ruminate block and retraining process makes LARA exceptionally suitable for online applications like web services monitoring.

The aforementioned approaches have their strengths and weaknesses, with every model excelling in specific types of datasets while also exhibiting limitations. For instance, the ML techniques have been foundational, but they often require extensive feature engineering and struggle with complex datasets (Chalapathy and Chawla, 2019). For DL approaches, reconstruction or forecasting-based models rely on reconstruction error to identify anomalies, they are more sensitive to large amplitude anomalies and may fail to detect subtle pattern differences or anomalies with small amplitude (Lee et al., 2023). In contrast, our proposed method can effectively capture anomalies with slight fluctuations by converting time series into images, and archive accurate few-shot detection result exploiting MLLMs' splendid generalization ability.

A.1.2 Time Series Anomaly Analysis.

Through a review of existing literature, we found that there is a lack of analysis on anomalies in current research. Common methods for analyzing anomalies identified by models involve visualizing the learned anomaly scores or parameters in relation to the ground truth (Dai and Chen, 2022; Lee et al., 2023), as well as taxonomy of the anomalies (Blázquez-García et al., 2021; Choi et al., 2021; Fahim and Sillitti, 2019). Yet, limited research has investigated the efficacy of proposed models in classifying different types of anomalies. For instance, (León-López et al., 2022) introduced a framework based on Hidden Markov Models for anomaly detection, supplemented by an additional supervised classifier to identify potential anomaly types. GIN (Wang and Liu, 2024) employs a two-stage algorithm that first detects anomalies using an informer-based framework enhanced with graph attention embedding, followed by classification of the detected anomalies through prototypical networks. Both aforementioned models rely on supervised training for their anomaly classification processes; consequently, the corresponding experiments conducted in these studies are limited to single classification datasets. In contrast, leveraging the capabilities of LLMs allows for not only the

identification of anomalous data points but also the provision of specific classifications and potential underlying causes for these anomalies, articulated in natural language and achieved in an unsupervised manner.

A.1.3 LLMs for time series.

Being pre-trained on enormous amounts of data, LLMs hold general knowledge that can be applied to numerous downstream tasks (Naveed et al., 2023; Min et al., 2023; Chang et al., 2024). Many researchers attempted to leverage the powerful generalization capabilities of LLMs to address challenges in time series tasks (Jin et al., 2023; Su et al., 2024; Li et al., 2024; Elhafsi et al., 2023). For instance, Gruver et al. (Gruver et al., 2024) developed a time series pre-processing scheme designed to align more effectively with the tokenizer used by LLMs. This approach can be illustrated as follows:

$$0.123, 1.23, 12.3, 123.0 \rightarrow \text{"1 2 , 1 2 3 , 1 2 3 0 , 1 2 3 0 0"}$$

Additionally, LSTPrompt (Liu et al., 2024) customizes prompts specifically for short-term and long-term forecasting tasks. Meanwhile, Time-LLM (Jin et al., 2023) reprograms input time series data using text prototypes and introduces the Prompt-as-Prefix (PaP) technique to further enhance the integration of textual and numerical information. Similarly, SIGLLM (Alnegheimish et al., 2024) is an LLM-based framework for anomaly detection with a moudle to convert time series data into language modality. Most of these efforts focus primarily on forecasting tasks and are largely confined to textual modalities.

Existing works remain constrained by the limited availability of sequential samples in the training datasets of LLMs (Merrill et al., 2024) and the models' inherent insensitivity to numerical data (Qian et al., 2022; Ye et al., 2024). Consequently, LLMs struggle to capture subtle changes in time series, making it difficult to produce reliable results (Merrill et al., 2024). While we recognize that natural language is a modality in which LLMs excel, it may not be the most effective format for processing time series data.

With the emergence of MLLMs (Zhang et al., 2024a), there is potential for enhanced reasoning capabilities that can accommodate a broader range of tasks beyond single-modal textual inputs (Wang et al., 2024a; Zhang et al., 2024c). Some research has indicated that these models possess analytical abilities for interpreting charts (Zhang et al., 2024b); however, no studies have yet applied them to the domain of anomaly detection in time series data. This gap highlights the need for further exploration into how MLLMs can be effectively utilized to detect and analyze anomalies based on visualized time series data.

A.2 Detailed Analysis

A.2.1 More on Large Multimodal Models

In this section, we attempt to explain the phenomenon of why the image modality in multimodal models appears to outperform the text modality for TSAD.

Feasibility: First, we emphasize the feasibility of the approach. The number of MLLMs is rapidly increasing (Yin et al., 2024), with open-source options such as Qwen-VL, LLaVA, and InternVL, as well as proprietary models like GPT-4o, Gemini, and Claude 3.5. These provide diverse and accessible choices for practitioners.

Intuitive Reasoning: Humans naturally perceive and interpret time series data through visual representations, such as plots, rather than by reading raw numerical values. Visualizations like line plots allow for immediate recognition of patterns, trends, and anomalies. Interestingly, certain characteristics of natural images align closely with time series data, such as smooth changes across most regions with abrupt transitions at edges (Chen et al., 2024b). This similarity reinforces the suitability of image-based approaches for representing and analyzing time series.

Theoretical Justification: Most MLLMs are pretrained on datasets that include tasks related to plot understanding, such as single-class and multi-class line plots (Methani et al., 2020; Masry et al., 2022). Since the visualizations of univariate time series essentially correspond to single-class line plots, these pretrained capabilities directly support the understanding of time series data (Wang et al., 2024b). Moreover, recent evidence highlights that specialized chart-related training data significantly enhances a model's ability to understand plots and charts (He et al., 2024).

Practical Insights: To further validate this, we analyzed the pretraining and post-training data (e.g., instruction tuning) of several MLLMs, including GPT-4o, Qwen-VL, and Gemini. These models incorporate datasets related to charts and plots,

such as ChartQA and academic articles or technical documents extracted from Common Crawl PDFs and HTML files[3]. This exposure provides them with robust chart/plot understanding capabilities.

To contrast, we experimented with open-source MLLMs such as LLaVA on TSAD tasks but observed significantly poorer performance. Upon investigation, we found that LLaVA's pretraining data lacked datasets related to plots or charts, which aligns with its weaker capabilities in handling time series visualizations.These findings collectively justify the use of MLLMs for TSAD tasks, as their pretrained knowledge on plots and charts directly aligns with the needs of analyzing time series data in visualized formats.

Instruction Tuning: Instruction tuning has the potential to enhance a model's adaptability. However, the primary focus of this work is to demonstrate that existing MLLMs, when integrated with our proposed TAMA framework, can effectively address TSAD tasks with robust interpretability. Due to constraints in time and computational resources, we did not pursue large-scale instruction tuning. Nonetheless, we believe this is a promising direction for future research, particularly when combined with efforts to enhance MLLMs' capabilities in understanding visual charts. We plan to explore this avenue further in subsequent work.

A.3 More on Experimental Setup

A.3.1 Dataset Details

Table 9: Details of all datasets. Datasets with classification labels include real-world datasets ($^+$) and a synthetic dataset (*) generated using GutenTAG (Wenig et al., 2022).

Dataset	#Train (K)	#Test (K) (labeled)	Anomaly% Point	Shapelet	Seasonal	Trend	Total
UCR$^+$	1.2-3.0	4.5-6.3	0.04	1.05	-	-	1.10
SMAP$^+$	0.3-2.9	4.5-8.6	-	7.0	0.2	0.1	7.3
MSL$^+$	0.4-4.3	1.1-6.1	1.3	6.2	-	3.0	10.5
NormA$^+$	-	104.0-196.0	-	18.6	4.1	1.2	24.0
Synthetic*	3.6	3.6	0.3	0	3.4	1.4	5.1
SMD	23.7-28.7	23.7-28.7	-	-	-	-	4.2
Dodgers	-	50.4	-	-	-	-	11.1
ECG	227.9-267.2	227.9-267.2	-	-	-	-	7.9

A.3.2 Some Suggestions about TAMA

In this paper, we propose a framework named TAMA to utilize the MLLM to analyze time series images. However, we have tried multiple versions and gained valuable practical experience during the development process. Based on our practical experience, we provide some suggestions.

- To better parse the output results, choosing the MLLM which supports JSON mode output or structured output can be very convenient. If the MLLM does not support these output format, we can use GPT-4o, which supports structured output, to format the output text.
- Assume the period of series data is T, it is recommended to set the sliding window length to at least $3T$.
- The MLLM marks the interval with anomaly based on the scale of the plot. Therefore, the scale of axis should be clear enough. However, the rotation of scale does not matter.
- Grid-like auxiliary lines can be added to enhance the accuracy of the anomaly intervals output by the MLLM.
- According to the documentation of OpenAI, in order to use high revolution mode, the figure size should not larger than 2000x768 pixels. All images in TAMA will be limited to this size.

A.3.3 The Usage of Tokens

Since the proposed framework, TAMA, utilizes MLLMs through the API calling, it is more meaningful to report the usage of API tokens rather than the model size. According to OpenAI's documentation [4], images are restricted to dimensions of pixels, with each image consuming up to 765 tokens. In TAMA, we will provide three reference images, one target image and two multi-scaled images if there is an anomaly detected in last stage. Therefore, the consumption of a normal case is $5 \times 765 = 3825$ tokens, the consumption of an abnormal case is $7 \times 765 = 5355$ tokens. In total, TAMA requires approximatly 7,000-8,000 tokens for comprehensive anlaysis of a single sample.

A.3.4 Prompts

The design of prompts is based on the documentation of OpenAI[5]. Writing the steps out explicitly can make it easier for the model to follow them. In our task, we separate the whole task into three specific tasks: **Multimodal Reference Learning** (see

[3] https://digitalcorpora.org/corpora/file-corpora/cc-main-2021-31-pdf-untruncated/

[4] https://platform.openai.com/docs/guides/

[5] https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions

Prompt 1), **multimodal Analyzing** (see Prompt 3) and **Multi-scaled Self-reflection** (see Prompt 2). Besides, we also provide some background information, such as siliding windows and additional information of images. With the JSON mode output of GPT-4o, it is very convenient for us to process the output results, requiring a detailed description of the output format in prompts. Based on our practical experience, we find that clear descriptions and a structured format significantly are very helpful for MLLM to understand.

A.4 Complete Experimental Results and Visualization

A.4.1 Full Results of Anomaly Detection across All Datasets

In this section, we present the full results of all datasets in Table 10. Due to the limitation of the space, we only present some of them in the main body. Meanwhile, we also present the variance in this table. Most of datasets contain more than one sub-sequence, to fully present and compare the performance, we evaluate all metrics in all subsequence and calculate three values: *mean*, *variance* and *maxima*. In this table, *mean* and *variance* are formated as "*mean* $\pm$ *variance*".

A.4.2 Full Results of the PAT Experiment

In Section 4.1, in order to study the impact of point-adjustment, we re-evaluate the results using the point-adjustment with a threshold α. Due to the limitation of the space, we only present results of some datasets in the main body. The full results are presented in Figure 6. As the figure presented, our framework achieves outstanding *AUC-PR* across all datasets at various α, showing that our framwork has better robustness and stability.

A.4.3 Full results of Anomaly Classification

Table 11 presents the type-specific anomaly detection performance. To maintain readability, only the F1-score without point adjustment is reported. The results highlight TAMA's outstanding performance in identifying pattern anomalies, including shapelet, seasonal, and trend types, while most baseline models struggle in this aspect without point adjustment. For instance, on the UCR-shapelet dataset, TAMA outperformed the second-best detector (GDN) by a substantial margin of 293% in terms of the mean F1-score. This superiority stems from TAMA's inherent ability to detect anomalous intervals. However, this characteristic may lead to lower F1-scores in the detection of point anomalies. In the synthetic dataset we generated, labels for point anomalies were strictly defined. While TAMA's interval detection always encompassed the ground-truth anomalies, it also produced a significant number of false positives.

A.4.4 Visualization of anomaly classification

In Section 4.2, we make a new dataset for anomaly classification by labeling some real-world datasets and generating sequence. We also provide some visualization of these anomalies to better understand the different types of anomalies. The visualization of anomaly classification is shown in Figure 7. The dataset contain four classification: Point, Shapelet, Seasonal and Trend, which are referenced from the work (Lai et al., 2021).

A.4.5 Case Studies of Abnormal Descriptions

In this section, we represent some case studies to show the interpretability of TAMA. The interpretability of TAMA refers to the ability to classify the anomaly type, describe the anomaly in detail and analyze possible causes based on the background information. We select four examples from UCR, ECG, SMD and Dodgers datasets, which are from different domains. The result of case studies is shown in Figure 8. In each case, the ground truth is markded in red and the detections of TAMA are marked in green. The detection results include the anomaly interval, the anomaly type and the anomaly description.

1. UCR datasets: The data we selected is named internal bleeding in UCR. The result is shown in (a) of Figure 8. In the anomaly description section, the interval (180, 200) is detected as a shapelet anomaly because TAMA finds this interval contains a series of rapid and sharp fluctuations that deviate from this smooth pattern. Based on the internal bleeding information background information, TAMA thinks this irregular patterns might be caused by some sudden changes in physilogical parameters, such as blood pressure or heart rate.

2. ECG dataset: This data records the electrocardiogram (ECG) data of ICU patients. The detection result on ECG dataset is shown in (b) of Figure 8. TAMA initially analyzes the frequency patterns in normal data and identifies that the interval between peaks occurring at indices 670 and 720 is substantially shorter than the expected periodicity, indicating a frequency anomaly. This deviation disrupts the regular periodic pattern typically

Table 10: Quantitative results across seven datasets use metrics point-adjusted *F1*, *AUC-PR*, and *AUC-ROC*. Best and second-best results are in bold and underlined, respectively. **TAMA** represents our framework, and **TAMA*** represents our framework without self-reflection. Each unit in the table contains two value: *mean* and *maxima* of all series. The number following the mean represents the standard deviation (*std*) computed over all sequences.

Dataset	UCR						NASA-SMAP						NASA-MSL					
Metric	F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%	
IF	24.7 ± 31.6	77.3	37.7 ± 15.0	44.6	24.4 ± 9.80	25.3	54.2 ± 36.4	94.2	58.9 ± 18.9	77.1	65.0 ± 6.10	87.7	47.6 ± 8.00	88.6	53.6 ± 4.80	80.4	68.7 ± 9.40	88.7
LOF	42.8 ± 1.10	**100**	35.6 ± 1.00	50.0	92.8 ± 19.2	**99.9**	62.2 ± 12.1	**100**	43.4 ± 12.9	61.4	60.1 ± 19.9	99.9	36.4 ± 25.3	66.8	44.5 ± 9.90	66.0	58.6 ± 0.50	99.8
TranAD	38.2 ± 40.7	93.7	30.9 ± 6.40	51.0	77.0 ± 26.8	**99.9**	59.0 ± 39.9	99.6	36.8 ± 19.1	73.9	74.4 ± 27.7	**100**	64.6 ± 38.6	99.1	49.2 ± 20.4	79.6	82.5 ± 18.3	99.9
GDN	71.4 ± 43.1	80.6	33.4 ± 0.40	59.0	87.1 ± 24.1	**99.9**	76.4 ± 38.5	**100**	40.8 ± 19.1	66.2	86.1 ± 27.7	**100**	85.1 ± 26.1	**100**	38.7 ± 9.90	56.7	93.8 ± 0.50	**100**
MAD_GAN	74.2 ± 40.4	85.0	51.5 ± 0.80	65.9	99.4 ± 1.30	**99.9**	61.3 ± 41.3	**100**	39.9 ± 19.1	72.3	83.3 ± 15.1	**100**	96.0 ± 5.60	**100**	46.4 ± 17.5	50.0	95.7 ± 7.10	**100**
MSCRED	32.6 ± 37.9	96.0	28.9 ± 2.70	45.9	94.2 ± 3.60	**99.9**	57.0 ± 44.2	97.9	40.8 ± 19.1	61.7	77.0 ± 28.2	**100**	63.0 ± 37.0	92.2	39.5 ± 16.3	51.8	73.2 ± 16.1	98.1
MTAD_GAT	14.8 ± 13.2	36.6	34.2 ± 4.80	38.9	84.6 ± 7.60	94.4	78.3 ± 37.7	**100**	40.2 ± 22.9	58.0	77.0 ± 12.3	**100**	90.6 ± 27.7	**100**	49.2 ± 9.70	67.8	81.2 ± 21.9	**100**
OmniAnomaly	34.5 ± 32.7	95.7	26.0 ± 0.30	45.9	85.6 ± 10.3	**99.9**	57.1 ± 39.9	**100**	43.6 ± 20.5	63.2	77.5 ± 24.8	**100**	71.4 ± 36.5	**100**	40.0 ± 17.2	74.9	85.0 ± 18.7	99.9
USAD	57.6 ± 35.6	**100**	33.1 ±0.40	50.0	97.1 ± 4.20	**99.9**	72.8 ± 35.8	**100**	43.6 ± 22.5	63.2	93.9 ± 9.10	**100**	91.6 ± 26.2	99.9	42.6 ± 9.90	60.8	94.2 ± 0.70	**100**
TimesNet	32.8 ± 8.30	45.8	15.4 ± 5.20	23.5	98.4 ± 1.10	99.4	**97.7** ± 3.50	**100**	51.4 ± 2.80	90.3	**99.8** ± 0.09	**100**	97.4 ± 4.70	**100**	52.9 ± 8.10	79.7	**99.8** ± 0.50	**100**
SIGLLM (GPT-4o)	23.1 ± 19.7	44.6	7.40 ± 6.70	15.5	93.5 ± 16.9	96.5	69.0 ± 34.4	97.8	29.1 ± 28.4	49.2	95.5 ± 3.60	99.8	70.7 ± 44.8	97.9	72.4 ± 28.6	**100**	90.0 ± 15.3	**100**
TAMA	**92.5** ± 17.9	97.6	**93.0** ± 12.1	**97.7**	**99.8** ± 0.10	**99.9**	94.5 ± 7.20	**100**	**95.5** ± 9.30	**100**	98.4 ± 4.60	**100**	**97.5** ± 2.10	**100**	**99.4** ± 17.8	**100**	**99.8** ± 0.20	**100**
TAMA*	**92.5** ± 17.9	97.6	**93.0** ± 12.1	**97.7**	**99.8** ± 0.10	**99.9**	87.8 ± 31.3	**100**	89.2 ± 16.6	**100**	97.0 ± 4.10	**100**	96.1± 4.30	**100**	97.7 ± 18.2	**100**	99.0 ± 0.20	**100**

Dataset	SMD						ECG						Dodgers					
Metric	F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%		F1%		AUC-PR%		AUC-ROC%	
IF	**83.9** ± 13.2	**100**	73.8 ± 17.3	97.0	99.5 ± 0.50	**100**	80.8 ± 20.5	99.0	73.4 ± 18.8	92.2	97.2 ± 4.70	**100**	48.4 ± 0.00	48.4	52.2 ± 0.00	52.2	**89.4** ± 0.00	**89.4**
LOF	27.8 ± 6.60	75.2	39.9 ± 1.90	64.6	52.9 ± 2.60	59.3	21.8 ± 12.0	39.8	41.4 ± 4.40	60.7	56.3 ± 10.8	84.2	45.3 ± 0.00	45.3	40.8 ± 0.00	40.8	63.0 ± 0.00	63.0
TranAD	77.0 ± 33.0	99.6	70.9 ± 31.9	91.0	96.8 ± 13.3	**100**	69.1 ± 23.2	98.9	74.7 ± 22.9	97.7	94.9 ± 6.30	**100**	38.2 ± 0.00	38.2	33.9 ± 0.00	33.9	74.6 ± 0.00	74.6
GDN	76.9 ± 0.90	99.7	55.0 ± 35.5	88.6	77.0 ± 1.60	**100**	75.2 ± 17.6	96.2	76.6 ± 17.5	97.4	96.9 ± 3.70	99.9	37.0 ± 0.00	37.0	31.3 ± 0.00	31.3	74.2 ± 0.00	74.2
MAD_GAN	67.1 ± 1.30	92.6	61.7 ± 37.2	87.5	91.7 ± 1.60	**100**	79.1 ± 19.5	99.3	79.2 ± 19.4	97.6	96.9 ± 3.80	**100**	32.2 ± 0.00	32.2	28.6 ± 0.00	28.6	74.7 ± 0.00	74.7
MSCRED	69.4 ± 30.2	95.7	55.8 ± 36.7	96.2	95.0 ± 14.5	**100**	66.4 ± 26.3	99.0	73.8 ± 22.6	97.2	88.7 ± 13.5	**100**	37.8 ± 0.00	37.8	30.6 ± 0.00	30.6	74.6 ± 0.00	74.6
MTAD_GAT	69.7 ± 18.5	95.3	59.5 ± 33.8	90.4	90.6 ± 6.00	99.7	67.5 ± 29.2	**100**	73.5 ± 23.8	**98.8**	82.5 ± 17.1	**100**	39.1 ± 0.00	39.1	36.0 ± 0.00	36.0	74.9 ± 0.00	74.9
OmniAnomaly	66.0 ± 6.80	96.4	61.6 ± 34.2	91.8	87.7 ± 2.60	**100**	76.8 ± 21.3	98.6	76.4 ± 24.0	97.5	93.5 ± 8.80	**100**	33.6 ± 0.00	33.6	35.4 ± 0.00	35.4	60.3 ± 0.00	60.3
USAD	72.2 ± 0.30	99.7	67.8 ± 35.6	93.5	94.4 ± 1.60	**100**	71.5 ± 20.8	96.9	75.2 ± 18.9	98.3	94.9 ± 5.90	**100**	37.8 ± 0.00	37.8	33.1 ± 0.00	33.1	74.6 ± 0.00	74.6
TimesNet	82.8 ± 25.3	**100**	57.3 ± 21.2	99.9	95.4 ± 11.3	**100**	**92.4** ± 3.70	96.6	**90.0** ± 5.60	97.6	**99.4** ± 0.40	**100**	48.1 ± 0.00	48.1	73.0 ± 0.00	73.0	83.7 ± 0.00	83.7
SIGLLM (GPT-4o)	42.9 ± 27.9	59.8	30.4 ± 21.0	53.1	68.8 ± 12.8	77.8	19.2 ± 13.7	50.4	71.0 ± 25.3	87.6	94.2 ± 3.20	96.9	48.1 ± 0.00	48.1	60.7 ± 0.00	60.7	83.2 ± 0.00	83.2
TAMA	77.8 ± 17.1	**100**	**87.9** ± 10.4	**100**	98.9 ± 1.40	**100**	81.3 ± 19.1	87.5	84.5 ± 15.4	90.0	95.4 ± 2.30	99.4	**65.6** ± 0.00	**65.6**	**74.0** ± 0.00	**74.0**	85.2 ± 0.00	85.2
TAMA*	62.8 ± 24.5	93.0	78.6 ± 14.1	97.2	**99.7** ± 1.50	99.7	78.1 ± 19.8	88.0	83.4 ± 14.6	91.1	94.7 ± 2.50	99.1	64.5 ± 0.00	64.5	73.6 ± 0.00	73.6	85.3 ± 0.00	85.3

Dataset	NormA					
Metric	F1%		AUC-PR%		AUC-ROC%	
IF	56.8 ± 19.2	86.3	52.3 ± 21.9	81.2	57.9 ± 1.00	68.7
LOF	54.5 ± 17.8	77.9	68.8 ± 9.30	92.4	95.1 ± 2.90	97.9
TranAD	38.0 ± 15.8	76.0	49.7 ± 21.3	78.9	53.6 ± 2.00	83.6
GDN	38.5 ± 14.8	74.7	50.9 ± 20.2	78.3	54.2 ± 2.10	82.2
MAD_GAN	38.5 ± 14.3	74.7	51.1 ± 19.8	77.8	54.1 ± 2.90	81.9
MSCRED	38.4 ± 16.1	74.6	49.7 ± 20.9	77.7	53.8 ± 2.00	81.8
MTAD_GAT	49.7 ± 13.7	93.8	50.1 ± 21.3	95.6	66.6 ± 3.30	94.2
OminiAnomaly	43.2 ± 17.9	74.8	53.5 ± 20.3	79.1	49.8 ± 1.70	89.9
USAD	38.6 ± 15.9	75.6	53.3 ± 20.9	78.5	54.1 ± 1.70	82.9
SIGLLM (GPT-4o)	82.8 ± 30.0	**94.6**	93.8 ± 21.4	**98.9**	97.9 ± 2.50	99.1
TAMA	80.7 ± 4.70	89.2	**95.0** ± 7.60	98.5	**98.1** ± 0.70	**99.2**
TAMA*	**83.9** ± 10.0	85.5	93.9 ± 10.8	98.7	97.4 ± 1.00	98.6

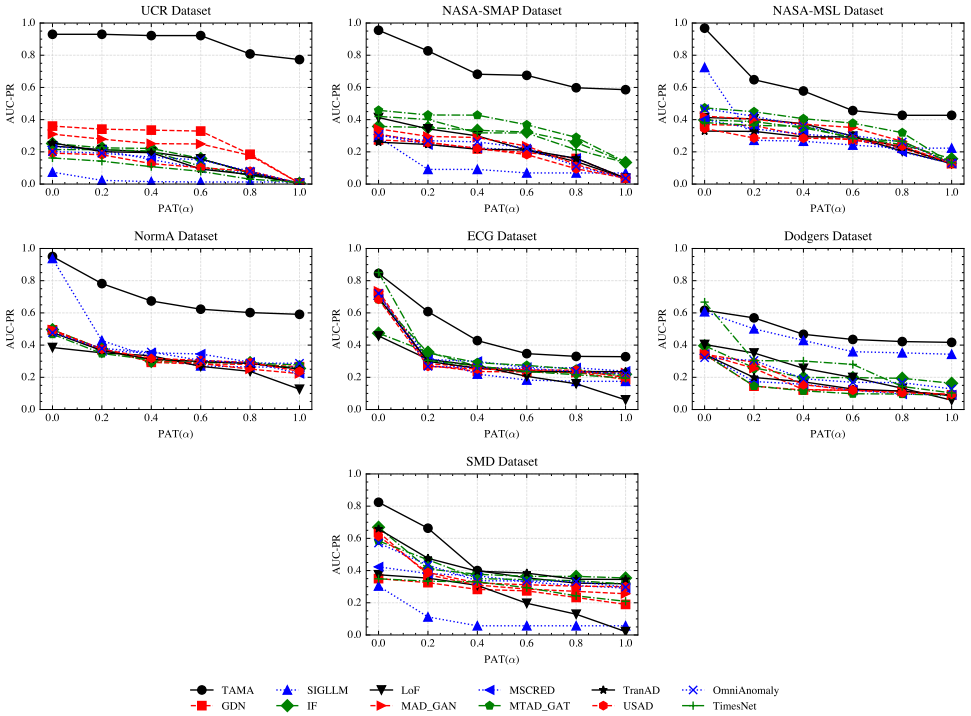


Figure 6: The full *AUC-PR* results of all models across all datasets at various point-adjustment threshold α (PAT, see Section 3.1).

Table 11: Quantitative results on each specific anomaly category across five datasets using F1-score% without point-adjustment. Best and second-best results are in bold and underlined, respectively. **TAMA** represents our framework (Some datasets include more than one series. To present the true performance of each method as much as possible, each unit in the table contains two values: *maxima* / *mean*. The *maxima* represents the best result among all sub-series, while the *mean* refers to the average of all sub-series.).

Dataset	**NASA-MSL**			**NASA-SMAP**			**UCR**	
Category	Point	Shapelet	Trend	Shapelet	Seasonal	Trend	Point	Shapelet
TranAD	23.2 / 10.4	14.9 / 13.5	33.7 / 33.7	1.00 / 0.60	1.40 / 1.40	0.30 / 0.30	0.30 / 0.30	10.5 / 3.60
GDN	3.80 / 2.30	17.6 / 8.60	1.20 / 1.20	6.30 / 2.10	1.20 / 1.20	0.30 / 0.30	**22.2 / 22.2**	53.0 / 20.6
MAD_GAN	3.90 / 2.00	17.6 / 8.50	1.60 / 1.60	10.2 / 3.00	1.20 / 1.20	0.65 / 0.65	14.0 / 14.0	36.8 / 15.0
MSCRED	61.9 / 23.0	16.6 / 8.00	23.4 / 23.4	1.00 / 0.60	0.70 / 0.70	0.30 / 0.30	0.45 / 0.45	2.40 / 0.80
MTAD_GAT	69.8 / **46.3**	16.6 / 8.00	**73.9 / 73.9**	52.0 / 24.0	2.10 / 2.10	0.95 / 0.95	0.80 / 0.80	2.70 / 1.45
OmniAnomaly	5.40 / 1.80	3.70 / 1.80	2.80 / 2.80	1.00 / 0.40	0.65 / 0.65	0.45 / 0.45	0.55 / 0.55	3.00 / 1.00
USAD	13.0 / 7.50	16.6 / 8.10	4.30 / 4.30	0.60 / 0.30	1.20 / 1.20	1.25 / 1.25	5.30 / 5.30	11.9 / 4.50
IF	35.0 / 24.2	30.0 / **22.9**	31.6 / 31.6	30.2 / 17.1	1.10 / 1.10	1.45 / 1.45	0.55 / 0.55	2.70 / 1.85
LoF	22.9 / 10.1	**33.4** / 22.4	33.3 / 33.3	14.0 / 7.80	0.60 / 0.60	4.60 / 4.60	0.45 / 0.45	2.05 / 1.35
TimesNet	23.2 / 10.9	12.5 / 8.30	22.4 / 10.8	20.4 / 8.95	1.35 / 1.35	25.6 / 25.6	0.40 / 0.40	1.80 / 1.20
SIGLLM	10.8 / 5.70	1.60 / 0.80	23.9 / 23.9	30.3 / 12.6	**20.2 / 20.2**	2.65 / 2.65	0.85 / 0.85	11.2 / 4.80
TAMA	**70.2** / 31.9	26.2 / 11.4	22.4 / 13.5	**77.4 / 47.9**	0.10 / 0.10	**84.5 / 84.5**	20.0 / 20.0	**92.3 / 81.0**

Dataset	**NormA**			**Synthetic**		
Category	Shapelet	Seasonal	Trend	Point	Seasonal	Trend
TranAD	4.00 / 2.30	3.30 / 2.20	3.90 / 2.50	0.55 / 0.35	8.90 / 1.90	13.5 / 7.90
GDN	4.10 / 2.30	3.30 / 2.20	3.90 / 2.50	0.50 / 0.35	13.6 / 2.10	12.4 / 6.50
MAD_GAN	4.10 / 2.30	3.30 / 2.20	3.90 / 2.50	0.55 / 0.35	8.50 / 1.60	13.9 / 8.00
MSCRED	4.10 / 2.30	3.30 / 2.20	3.90 / 2.50	0.55 / 0.35	8.50 / 1.60	10.5 / 5.20
MTAD_GAT	4.90 / 1.40	1.30 / 0.90	1.70 / 1.10	0.55 / 0.40	6.30 / 1.80	13.3 / 6.25
OmniAnomaly	12.6 / 5.70	11.5 / 7.50	11.4 / 7.30	30.3 / 19.8	9.30 / 1.70	11.2 / 7.80
USAD	4.10 / 2.40	3.30 / 2.20	3.90 / 2.50	0.55 / 0.35	8.90 / 1.70	12.6 / 7.30
IF	21.4 / 13.2	17.0 / 12.9	13.9 / 11.9	36.2 / 21.6	10.4 / 9.05	10.9 / 7.20
LoF	30.7 / 16.8	25.7 / 18.6	21.5 / 16.9	0.50 / 0.50	10.9 / 9.10	5.35 / 5.25
TimesNet	10.2 / 9.05	5.25 / 5.20	1.79 / 1.60	**37.5 / 25.5**	11.9 / 9.50	10.6 / 5.70
SIGLLM	6.50 / 3.10	3.70 / 2.50	0.90 / 0.60	0.60 / 0.35	10.9 / 7.95	14.0 / 6.50
TAMA	**56.8 / 37.1**	**38.8 / 28.1**	**45.2 / 34.3**	3.90 / 1.80	**27.1 / 18.4**	**14.1 / 8.20**

observed in normal data. Furthermore, based on contextual analysis, TAMA suggests two potential underlying causes: cardiac ischemia or arrhythmia.

3. SMD dataset: The Server Machine Dataset (SMD) is a comprehensive multivariate time series dataset collected from 28 different servers at a large Internet company over a continuous five-week period. Each server records 38-dimensional metrics at one-minute intervals, making it particularly valuable for anomaly detection research. As the (c) in Figure 8 shown, TAMA detects correctly two anomaly intervals from three. The description section explains the reason of this detection. TAMA frist reads the values of time series data. Compared with the normal data, the peaks in invertals (900, 1100) and (3500, 3700) are significantly higher than others, which are considered as the global anomaly.

TAMA, however, failed to detect one anomaly in the 1600-2000 interval, which can be attributed to two primary factors. First, the pronounced cyclical patterns in the data diminish the distinction between normal and abnormal patterns. Second, despite the data containing approximately three cycles, each cycle is compromised by anomalies, resulting in imprecise estimation of the baseline pattern. Based on background information analysis, three key factors potentially contribute to these anomalies. First, there is an abrupt surge in server activity and resource utilization. Second, system malfunction appears to be a significant contributing factor. Third, the anomalies may be attributed to external security breaches or cyber attacks.

4. Dodgers dataset: This data was collected for the Glendale on ramp for the 101 North freeway in Los Angeles. It is close enough to the stadium to see unusual traffic after a Dodgers game, but not so close and heavily used by game traffic so that the signal for the extra traffic is overly obvious. The observations were taken over 25 weeks, 288 time slices per day (5 minute count aggregates). The result is shown in (d) of Figure 8. TAMA marks two contextual anomalies in this case. TAMA can recognise the 'contextual' anomaly between 570

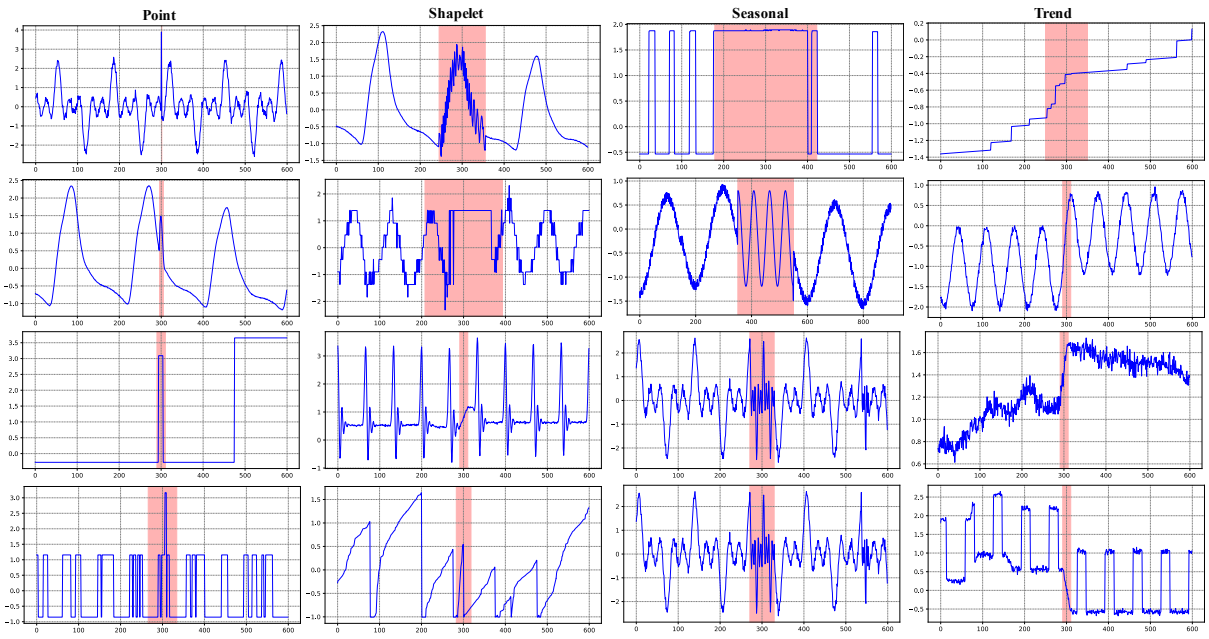


Figure 7: Visualization of anomalies. Each row displays sequences from different datasets that contain the same type of anomaly.

and 600 correctly. Based on the background information, TAMA thinks the first sharp spike anomaly could represent a sudden increase in traffic as people leave the stadium. The second anomaly, a sudden drop to a flat line, possibly due to sensor malfunction or maintenance activities.

The whole inference of TAMA reveals that the MLLM has the ability of understanding anomaly and make a reasonable detection. Moreover, TAMA can try to inference the possible causes based on the background information given in the prompt.

A.4.6 Ablation Study for Other Factors

We also conduct some ablation experiments to evaluate the impact of each factor.

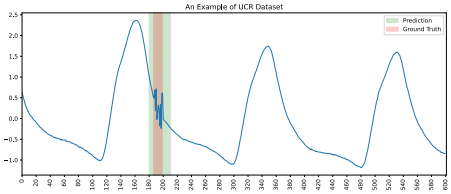
Window Size. In TAMA, as it shown before (in Section 3.1), we use sliding window in the data pre-processing stage. To accommodate different time series data with varying periods, we report the window size in multiples of the data period.

The results presented in Figure 9 reveal that the performance of our method is positively correlated with the window size. This is because the MLLMs struggles to identify periodic patterns when given only single-period images, resulting in incorrectly classifying periodic features and truncated features as anomalies. Therefore, we ultimately set the window size to approximately $3T$ for the experiments detailed in Section 4.1.

Additional Information in Images. The transformation of raw data into visual formats, such as images, adds crucial information, including plot orientation and auxiliary lines. This study investigates how these elements influence TAMA's performance in identifying abnormal intervals based on plot scales. We conducted two experiments: the first involved rotating images by 90 degrees before inputting them into TAMA, while the second examined the impact of auxiliary lines, which are perpendicular to the x-axis and align with the scale to aid in locating data points.

Both experiments are performed on the UCR and NASA-SMAP datasets. Results are presented in Table 12, where TAMA represents the original model, TAMA-R indicates performance with rotated images, and TAMA-A reflects performance without auxiliary lines. We evaluated using the AUC-PR without point adjustment. The findings demonstrate a notable decline in TAMA-R's performance with rotated images, suggesting that the MLLMs are sensitive to image orientation. Despite the rotation of axis is disturbed in prompts, the MLLM struggles to interpret rotated images accurately, leading to reduced anomaly detection. In contrast, TAMA-A experiences only a slight performance decrease across both datasets, indicating that MLLMs can better identify abnormal intervals when auxiliary lines are present.

These experiments reveal that MLLMs perceive



Detection :

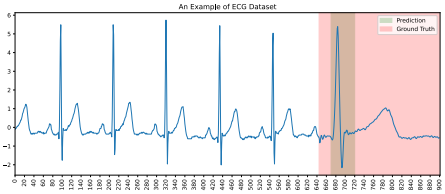
- **Anomaly range**: (180, 210)
- **Anomaly type**: shapelet

Anomaly Description:

The abnormality observed between indices 180 and 210 is best classified as a 'shapelet' outlier. Shapelet outliers refer to subsequences with dissimilar basic shapelets compared to the normal shapelet. In this case, the normal shapelet is a smooth and regular oscillation with consistent peaks and troughs. The abnormal section, however, shows a series of rapid and sharp fluctuations that deviate from this smooth pattern.

The abnormality in the time series data, characterized by rapid oscillations and fluctuations, could potentially be indicative of an internal bleeding event. In medical monitoring, such irregular patterns might reflect sudden changes in physiological parameters, such as blood pressure or heart rate, which can occur during episodes of internal bleeding. The disruption in the otherwise stable pattern suggests an acute event that requires further investigation to confirm the presence and cause of internal bleeding.

(a) UCR



Detection :

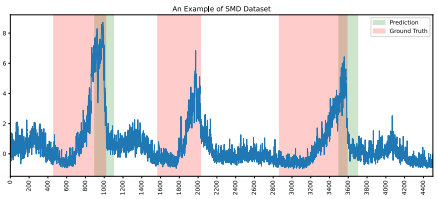
- **Anomaly interval**: (670. 720)
- **Anomaly type**: frequency

Anomaly Description:

The abnormality detected in the time series data slice is classified as a 'frequency' anomaly. Frequency outliers refer to changes in the regularity or periodicity of the data. In the normal reference data, the peaks occur at regular intervals of approximately 120 time units. However, in this data slice, the interval between the peaks around indices 670 and 720 is significantly shorter than the expected interval. This change in the frequency of the peaks disrupts the regular periodic pattern observed in the normal data.

The abnormality observed in the ECG data, characterized by a shift in the pattern and a lack of return to the baseline, could be indicative of a physiological change or issue. One possible cause could be a cardiac event such as ischemia, where the heart muscle receives insufficient blood flow, leading to changes in the ECG pattern. Another possibility is an arrhythmia, where the heart beats irregularly, disrupting the normal periodic pattern. Additionally, external factors such as electrode displacement or interference from other electronic devices could also cause such abnormalities in the ECG data. It is important to consult a medical professional for an accurate diagnosis.

(b) ECG



Detection :

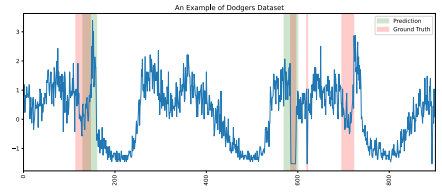
- **Anomaly interval**: (900, 1100), (3500, 3700)
- **Anomaly type**: global, global

Anomaly Description:

The interval (900, 1100) shows a significant deviation from the normal pattern, with an unusually high peak reaching approximately 8, which is much higher than the typical peaks observed in the normal reference data slices. This peak is sharp and prominent, standing out clearly from the surrounding data points, making it a global outlier. I also identified another abnormality between indices 3500 and 3700. This interval also shows a significant deviation from the normal pattern, with an unusually high peak reaching approximately 7, which is much higher than the typical peaks observed in the normal reference data slices. This peak is sharp and prominent, standing out clearly from the surrounding data points, making it another global outlier.

The abnormalities observed in the time series data, characterized by sharp and unusually high peaks, could be attributed to sudden spikes in server activity or resource usage. These spikes might be caused by unexpected workload increases, such as a surge in user requests or data processing tasks. Other potential causes could include system malfunctions, such as hardware failures or software bugs, leading to erratic behavior in the server's performance metrics. Additionally, external factors like network issues or security breaches could also result in such anomalies. Given the context of server monitoring, these anomalies likely indicate periods where the server's performance deviated significantly from its normal operating conditions, warranting further investigation to identify and address the underlying issues.

(c) SMD



Detection :

- **Anomaly interval**: (130, 160), (570, 600)
- **Anomaly type**: contextual, contextual

Anomaly Description:

The first abnormality, occurring between indices 570 and 600, is classified as a 'contextual' outlier. This is because the sudden drop and subsequent recovery in the data values deviate significantly from the surrounding context, which is characterized by relatively stable fluctuations. The abrupt nature of this drop and recovery is not consistent with the smooth transitions observed in the normal reference data, making it a clear contextual anomaly. The second abnormality, observed between indices 130 and 160, is also classified as a 'contextual' outlier. The unusual spike in the data values during this interval deviates from the expected pattern of smooth transitions between peaks and troughs. This spike is abrupt and significantly higher than the typical peaks in the normal reference data, indicating a contextual anomaly. Both abnormalities are identified based on their deviation from the expected periodic and stable pattern observed in the normal reference data slices.

The abnormalities in the time series data likely correspond to traffic patterns influenced by events at Dodgers Stadium. The first abnormality, a sharp spike between indices 130 and 170, could represent a sudden increase in traffic as people leave the stadium after a game. This aligns with the note that event traffic begins at or near the end of the event time. The second abnormality, a sudden drop to a flat line between indices 580 and 620, might indicate a temporary disruption or error in data collection, possibly due to sensor malfunction or maintenance activities. These patterns suggest that the observed anomalies are related to the impact of baseball games on local traffic flow.

(d) Dodgers

Figure 8: Case studies of abnormal descriptions.

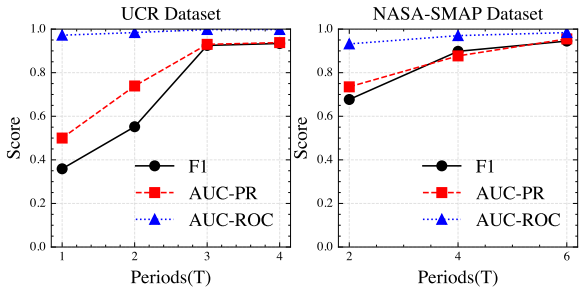


Figure 9: Results of window size ablation experiments. For the period of two datasets, $T_{UCR} \approx 200$, $T_{NASA-SMAP} \approx 100$

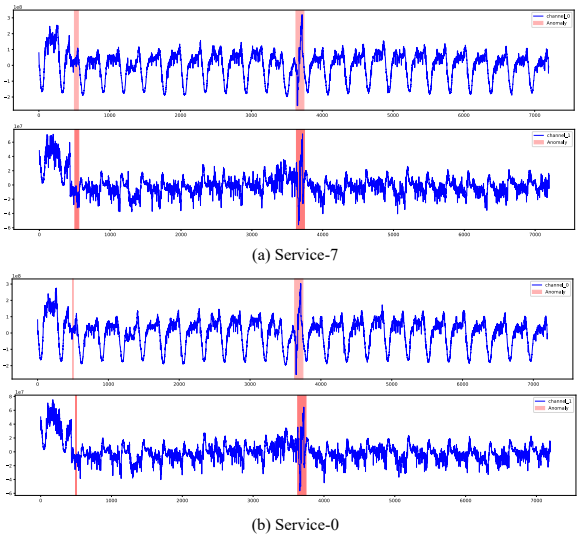
time series images similarly to humans— uxiliary lines enhance anomaly localization accuracy, while image rotation negatively affects performance. This sensitivity may result from the tokenizer's responsiveness to orientation or insufficient training data and guidance.

Table 12: The average *AUC-PR%* performance of TAMA with different additional information in images.

Datasets	TAMA	TAMA-R	TAMA-A
UCR	83.0	32.9 (**-50.1**)	75.6 (**-7.60**)
NASA-SMAP	72.9	28.6 (**-44.3**)	66.4 (**-6.50**)

A.4.7 PCA Dimensionality Reduction Visualization

The visualization of SMD data after dimensionality reduction using PCA is shown in Figure 10. Since PCA retains the components with the highest variance, shapelet anomalies are often preserved effectively. However, for certain seasonal anomalies, their characteristics may become less pronounced, which reflects some of the information loss associated with PCA.



(a) Service-7

(b) Service-0

Figure 10: The visualization of SMD data after dimensionality reduction using PCA. (a) and (b) are collected from Service-7 and Service-0 in the SMD dataset respectively.

A.5 License

Table 13: MLLMs Involved in Experiments and Their Corresponding Licenses

Model Name	**License Type**
GPT-4o (-mini)	Terms of Use
Gemini-1.5 (-pro/-flash)	Apache License 2.0
Qwen-VL-Max-0809	Apache License 2.0

Prompt 1: Multimodal Reference Learning Prompt

<Background>: I have a long time series data with some abnormalities. I have converted the data into plots and I need your help to find the abnormality in the time series data. This task contains two parts:

- "Task1": I will give you some "normal reference" time series data slices without any abnormality. And you need to extract some valuable information from them to help me find the abnormality in the following time series data slices.
- "Task2": I will give you some time series data slices with some abnormalities. You need to find the abnormality in them and provide some structured information.

besides, I will offer you some background information about the data plots:

- The horizontal axis represents the time series index.
- The vertical axis represents the value of the time series.
- all normal reference data slices are from the same data channel but in different strides. Therefore, some patterns based on the position, for example, the position of peaks and the end of the plot, may cause some confusion.
- all normal references are slices of the time series data with a fixed length and the same data channel. Therefore the beginning and the end of the plot may be different but the pattern should be similar.

<Task>: Now we are in the "Task1" part: I will give you some "normal reference" time series data slices without any abnormality. And you need to extract some valuable information from them to help me find the abnormality in the following time series data slices.
<Target>: Please help me extract some valuable information from them to help me find the abnormality in the following time series data slices. The output should include some structured information, please output in JSON format:

- normal_pattern (a 300-400 words paragraph): Try to describe the pattern of all "normal references" . All normal reference data slices are from the same data channel but in different strides. The abnormal pattern caused by truncation might be found at the beginning and end of the sequence, do not pay too much attention to them. The description should cover at least the following aspects: period, stability, trend, peak, trough, and other important features.

Prompt 2: Multi-scaled Self-reflection

<Background>: I have a long time series data with some abnormalities. I have converted the data into plots and I need your help to find the abnormality in the time series data. There has been a response from another assistant, but I am not sure about the prediction. I need your help to double check the prediction. Besides, I will offer you some background information about the data plots:

- The horizontal axis represents the time series index.
- The vertical axis represents the value of the time series.
- all normal reference data slices are from the same data channel but in different strides. Therefore, some patterns based on the position, for example, the position of peaks and the end of the plot, may cause some confusion.
- all normal references are slices of the time series data with a fixed length and the same data channel. Therefore, the beginning and the end of the plot may be different, but the pattern should be similar.

<Task>: Now, I will give you some "normal reference" and you are expected to double check the prediction of the abnormality in the given data.
<Target>: The prediction of another assistant contains some information as follows:

- abnormal_index: The abnormality index of the time series. The output format should be like "[(start1, end1)/confidence_1/abnormal_type_1, (start2, end2)/confidence_2/abnormal_type_2, ...]", if there are some single outliers, the output should be "[(index1)/confidence_1/abnormal_type_1, (index2)/confidence_2/abnormal_type_2, ...]",if there is no abnormality, you can say "[]".
- abnormal_description: Make a brief description of the abnormality, why do you think it is abnormal?

Based on the "nomral reference" I gave you, please read the prediction above and double check the prediction. If you find any mistakes, please correct them. The output should include some structured information, please output in JSON format:

- corrected_abnormal_index (string, the output format should be like "[(start1, end1)/confidence_1/abnormal_type_1, (start2, end2)/confidence_2/abnormal_type_2, ...]", if there are some single outliers, the output should be "[(index1)/confidence_1/abnormal_type_1, (index2)/confidence_2/abnormal_type_2, ...]",if there is no abnormality, you can say "[]". The final output should be mixed with these three formats.): The abnormality index of the time series. There are some requirements:
  + 1. you should check each prediction of the abnormal_type and make sure it is correct based on the abnormality index. If there is a incorrect prediction, you should remove it.
  + 2. you should check each prediction of the abnormal_index according to the image I gave to you. If there is an abnormality in image but not in the prediction, you should add it. The format should keep the same as the original prediction.
- The reason why you think the prediction is correct or incorrect. (a 200-300 words paragraph): Make a brief description of your double check, why do you think the prediction is correct or incorrect?

Prompt 3: Multimodal Analyzing Prompt

<Background>: I have a long time series data with some abnormalities. I have converted the data into plots and I need your help to find the abnormality in the time series data. This task contains two parts:

- "Task1": I will give you some "normal reference" time series data slices without any abnormality. And you need to extrace some valuable information from them to help me find the abnormality in the following time series data slices.

- "Task2": I will give you some time series data slices with some abnormalities. You need to find the abnormality in them and provide some structured information.

Besides, I will offer you some background information about the data plots:

- The horizontal axis represents the time series index.

- The vertical axis represents the value of the time series.

- all normal reference data slices are from the same data channel but in different strides. Therefore, some patterns based on the position, for example, the position of peaks and the end of the plot, may cause some confusion.

- all normal references are slices of the time series data with a fixed length and the same data channel. Therefore the beginning and the end of the plot may be different but the pattern should be similar.

<Task>: In "Task1" part, you have already extracted some valuable information from the "normal reference" time series data slices. You can use them to help you find the abnormality in the following time series data slices. Now we are in "Task2", you are expected to detect the abnormality in the given data.
<Target>: Please help me find the abnormality in this time series data slice and provide some structured information. The output should include some structured information, please output in JSON format:

- abnormal_index (the output format should be like "[(start1, end1)/confidence_1/abnormal_type_1, (start2, end2)/confidence_2/abnormal_type_2, ...]", if there is no abnormality, you can say "[]". The final output should be mixed with these three formats.): The abnormality index of the time series. There are some requirements:
    + There may be multiple abnormalities in one stride. Please try to find all of them. Pay attention to the range of each abnormality, the range should cover each whole abnormality in a suitable range.
    + Since the x-axis in the image only provides a limited number of tick marks, in order to improve the accuracy of your prediction, please try to estimate the coordinates of any anomaly locations based on the tick marks shown in the image as best as possible.
    + all normal reference data slices are from the same data channel but in different strides. Therefore, some patterns based on the position, for example, the position of peaks and the end of the plot, may cause some confusion.
    + abnormal_type (answer from "global", "contextual", "frequency", "trend", "shapelet"): The abnormality type of the time series, choose from [global, contextual, frequency, trend, shapelet]. The detailed explanation is as follows:
        + global: Global outliers refer to the points that significantly deviate from the rest of the points. Try to position the outliers at the center of the interval.
        + contextual: Contextual outliers are the points that deviate from its corresponding context, which is defined as the neighboring time points within certain ranges. Try to position the outliers at the center of the interval.
        + frequency: Frequency outliers refer to changes in frequency, the basic shape of series remains the same. Please focuse on the horizontal axis to find the frequency anomalies.
        + trend: Trend outliers indicate the subsequences that significantly alter the trend of the time series, leading to a permanent shift on the mean of the data. Mark the intervals where the mean of the data significantly changes.
        + shapelet: Shapelet outliers refer to the subsequences with totally different shapes compared to the rest of the time series.

- confidence (integer, from 1 to 4): The confidence of your prediction. The value should be an integer between 1 and 4, which represents the confidence level of your prediction. Each level of confidence is explained as follows:
    + 1: No confidence: I am not sure about my prediction
    + 2: Low confidence: Weak evidence supports my prediction
    + 3: medium confidence: strong evidence supports my prediction
    + 4: high confidence: more than 95
    + based on the provided abnormal_type, you should double check the abnormal_index.

- abnormal_description (a 200-300 words paragraph): Make a brief description of the abnormality, why do you think it is abnormal?

- abnormal_type_description (a 200-300 words paragraph): Make a brief description of the abnormality type for each prediction, why do you think this type is suitable for the abnormality?

Last, please double check before you submit your answer.