

Detect, Explain, Interpret: an end-to-end benchmark for Time Series Anomaly Detection, Explainability, and Interpretability

Roberto Stanzione, Jules Barbe, Magali Parrino, Jérémie Fourmann, Paul Boniol

ENS UNIVERSITÉ PARIS SCIENCES & LETTRES

Inria

SCALITY

40th Annual Conference on Neural Information Processing Systems *Sydney, December 6-12, 2026*

Introduction

- Time Series **Anomaly Detection (TSAD)** aims to identify patterns that **deviate from normal behaviour** [1]

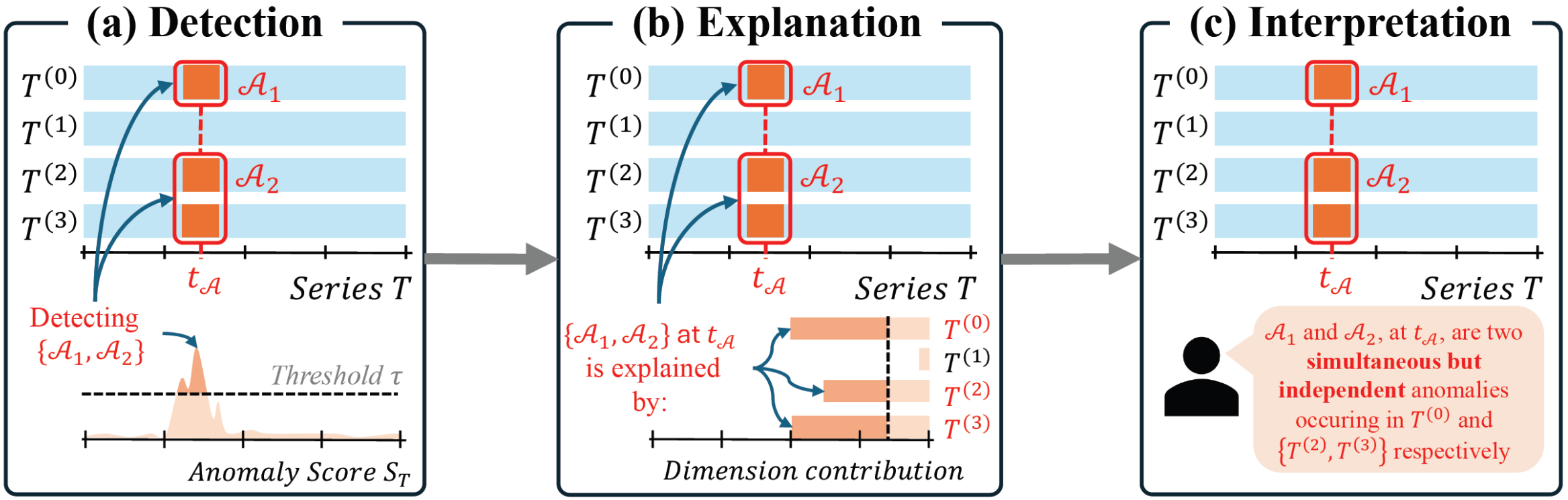
- The increasing attention on TSAD led to the **introduction of several benchmarks** to compare detectors

- However, these **only focus on detection**, overlooking **explainability** and **interpretability**

Benchmark	Dataset			Evaluation		
	Multi.	Ann.	Desc.	AD	Expl.	Interp.
NAB [50]	×	×	×	✓	×	×
Exathlon [38]	✓	×	×	✓	✓	×
TODS [46]	✓	×	×	✓	×	×
TimeEval [75]	✓	×	×	✓	×	×
TSB-UAD [67]	×	×	×	✓	×	×
TimeSeAD [87]	✓	×	×	✓	×	×
Exp-TSAD [103]	✓	×	×	✓	×	×
TSB-AD [54]	✓	×	×	✓	×	×
TAB [68]	✓	×	×	✓	×	×

Ann.: per-dimension labels **Desc.:** per-dimension description
AD: evaluation of anomaly detection performance

Research Gap



Per-timestamp labels

Benchmarks provide a global timestamp label to evaluate temporal anomaly localization

No fine-grained labels

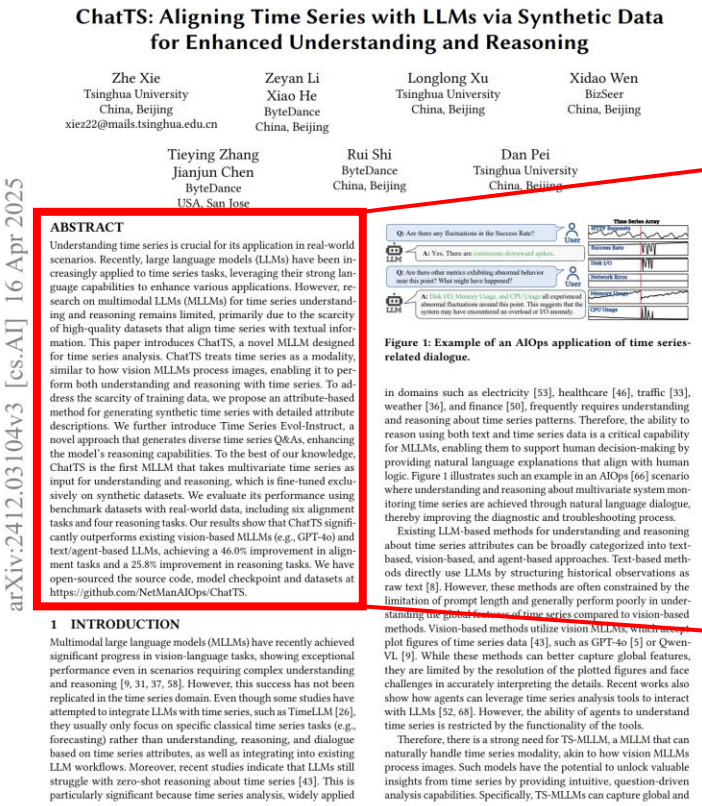
Only binary labels per time step: no affected dimension, no nature of the anomaly.

No contextual knowledge

No semantic annotations, so human-readable interpretations cannot be evaluated.

Research Gap

- The **lack** of benchmarks with **contextual knowledge** is also **limiting** the research for **time series analysis** and **reasoning** [2]:



ChatTS: Aligning Time Series with LLMs via Synthetic Data for Enhanced Understanding and Reasoning

Zhe Xie, Tsinghua University, China, Beijing, xiez22@mails.tsinghua.edu.cn
Zeyan Li, Xiao He, ByteDance, China, Beijing
Longlong Xu, Tsinghua University, China, Beijing
Xidao Wen, BizSeer, China, Beijing
Tieying Zhang, Jianjun Chen, ByteDance, USA, San Jose
Rui Shi, ByteDance, China, Beijing
Dan Pei, Tsinghua University, China, Beijing

arXiv:2412.03104v3 [cs.AI] 16 Apr 2025

ABSTRACT

Understanding time series is crucial for its application in real-world scenarios. Recently, large language models (LLMs) have been increasingly applied to time series tasks, leveraging their strong language capabilities to enhance various applications. However, research on multimodal LLMs (MLLMs) for time series understanding and reasoning remains limited, primarily due to the scarcity of high-quality datasets that align time series with textual information. This paper introduces ChatTS, a novel MLLM designed for time series analysis. ChatTS treats time series as a modality, similar to how vision MLLMs process images, enabling it to perform both understanding and reasoning with time series. To address the scarcity of training data, we propose an attribute-based method for generating synthetic time series with detailed attribute descriptions. We further introduce Time Series Evol-Instruct, a novel approach that generates diverse time series Q&As, enhancing the model's reasoning capabilities. To the best of our knowledge, ChatTS is the first MLLM that takes multivariate time series as input for understanding and reasoning, which is fine-tuned exclusively on synthetic datasets. We evaluate its performance using benchmark datasets with real-world data, including six alignment tasks and four reasoning tasks. Our results show that ChatTS significantly outperforms existing vision-based MLLMs (e.g., GPT-4o) and text/agent-based LLMs, achieving a 46.0% improvement in alignment tasks and a 25.8% improvement in reasoning tasks. We have open-sourced the source code, model checkpoint and datasets at https://github.com/NetManAIOps/ChatTS.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently achieved significant progress in vision-language tasks, showing exceptional performance even in scenarios requiring complex understanding and reasoning [9, 31, 37, 58]. However, this success has not been replicated in the time series domain. Even though some studies have attempted to integrate LLMs with time series, such as TimeLLM [26], they usually only focus on specific classical time series tasks (e.g., forecasting) rather than understanding, reasoning, and dialogue based on time series attributes, as well as integrating into existing LLM workflows. Moreover, recent studies indicate that LLMs still struggle with zero-shot reasoning about time series [43]. This is particularly significant because time series analysis, widely applied

Figure 1: Example of an AIOps application of time series-related dialogue.

in domains such as electricity [53], healthcare [46], traffic [33], weather [36], and finance [50], frequently requires understanding and reasoning about time series patterns. Therefore, the ability to reason using both text and time series data is a critical capability for MLLMs, enabling them to support human decision-making by providing natural language explanations that align with human logic. Figure 1 illustrates such an example in an AIOps [66] scenario where understanding and reasoning about multivariate system monitoring time series are achieved through natural language dialogue, thereby improving the diagnostic and troubleshooting process.

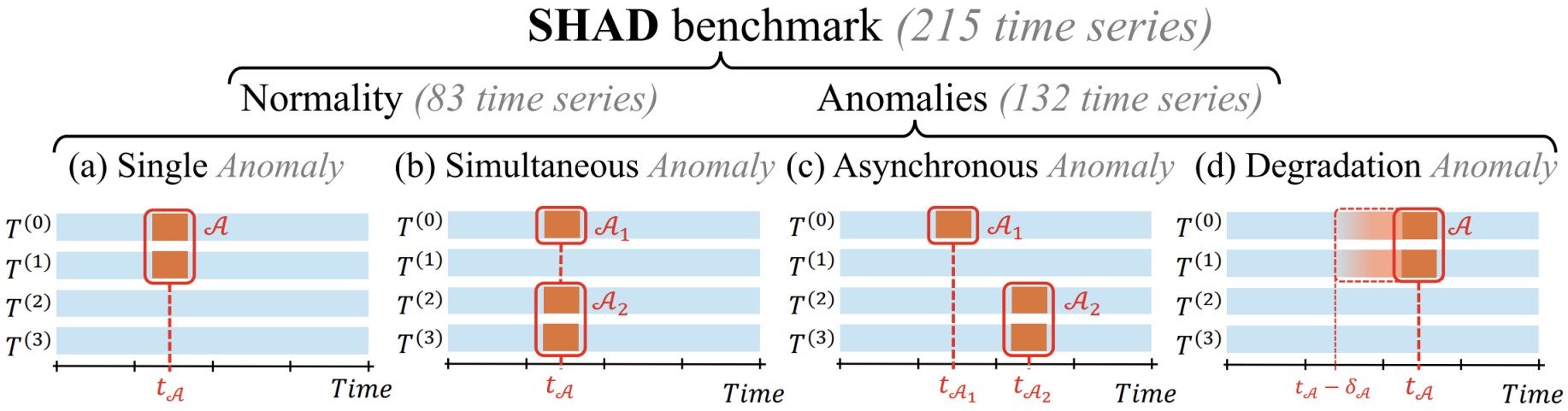
Existing LLM-based methods for understanding and reasoning about time series attributes can be broadly categorized into text-based, vision-based, and agent-based approaches. Text-based methods directly use LLMs by structuring historical observations as raw text [8]. However, these methods are often constrained by the limitation of prompt length and generally perform poorly in understanding the global features of time series compared to vision-based methods. Vision-based methods utilize vision MLLMs, which accept plot figures of time series data [43], such as GPT-4o [5] or Qwen-VL [9]. While these methods can better capture global features, they are limited by the resolution of the plotted figures and face challenges in accurately interpreting the details. Recent works also show how agents can leverage time series analysis tools to interact with LLMs [52, 68]. However, the ability of agents to understand time series is restricted by the functionality of the tools.

Therefore, there is a strong need for TS-MLLM, a MLLM that can naturally handle time series modality, akin to how vision MLLMs process images. Such models have the potential to unlock valuable insights from time series by providing intuitive, question-driven analysis capabilities. Specifically, TS-MLLMs can capture global and

> “However, **research […] for time series understanding and reasoning remains limited**, primarily **due to the scarcity of high-quality datasets that align time series with textual information**.”

SHAD: a new benchmark

- We present **SHAD** (**S**cality **H**igh-Dimensional **A**nomaly **D**etection benchmark), a benchmark that supports the evaluation of a complete TSAD pipeline
- SHAD is built **in collaboration with Scality**, deploying an experimental framework on the **RING distributed object storage system**



Disk	✓ (10 time series)	✓ (17 time series)	✓ (11 time series)	✗ (0 time series)
Server	✓ (11 time series)	✗ (0 time series)	✓ (11 time series)	✓ (26 time series)
Storage Node	✓ (16 time series)	✓ (14 time series)	✓ (16 time series)	✗ (0 time series)

171
Dimensions per series

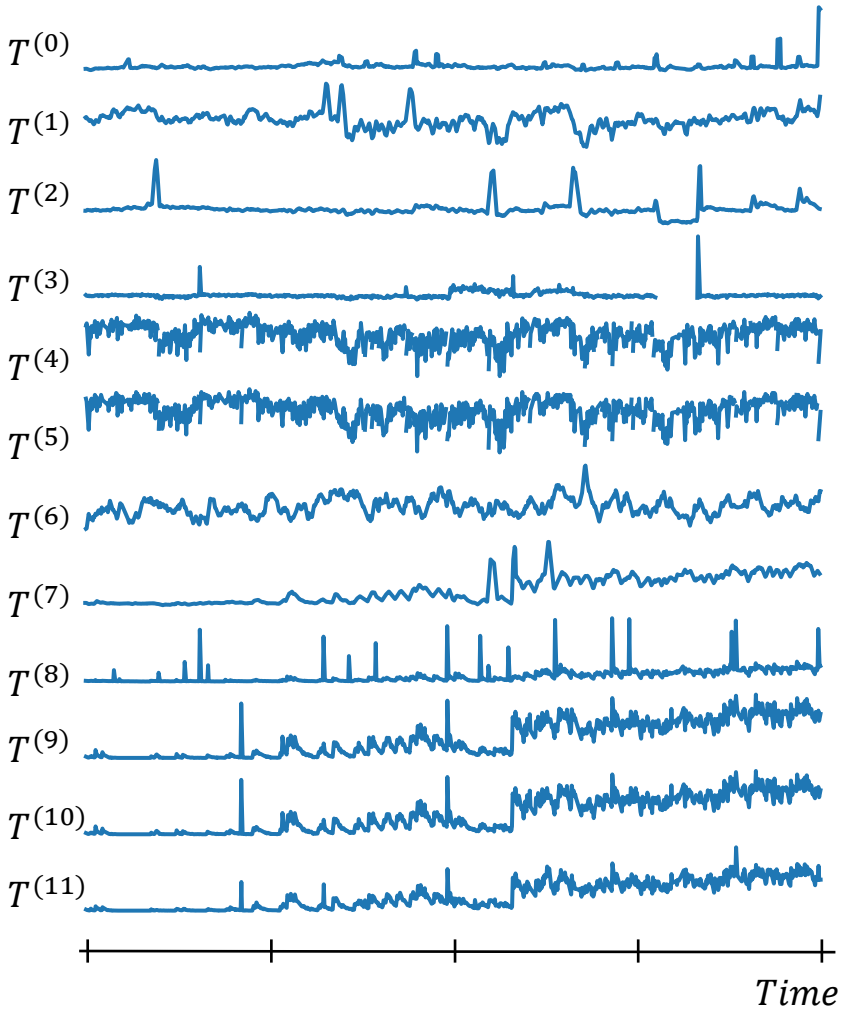
215
High-Dimensional series

83 / 132
Nominal / Anomalous series

18 h
Per run, 1-minute sampling

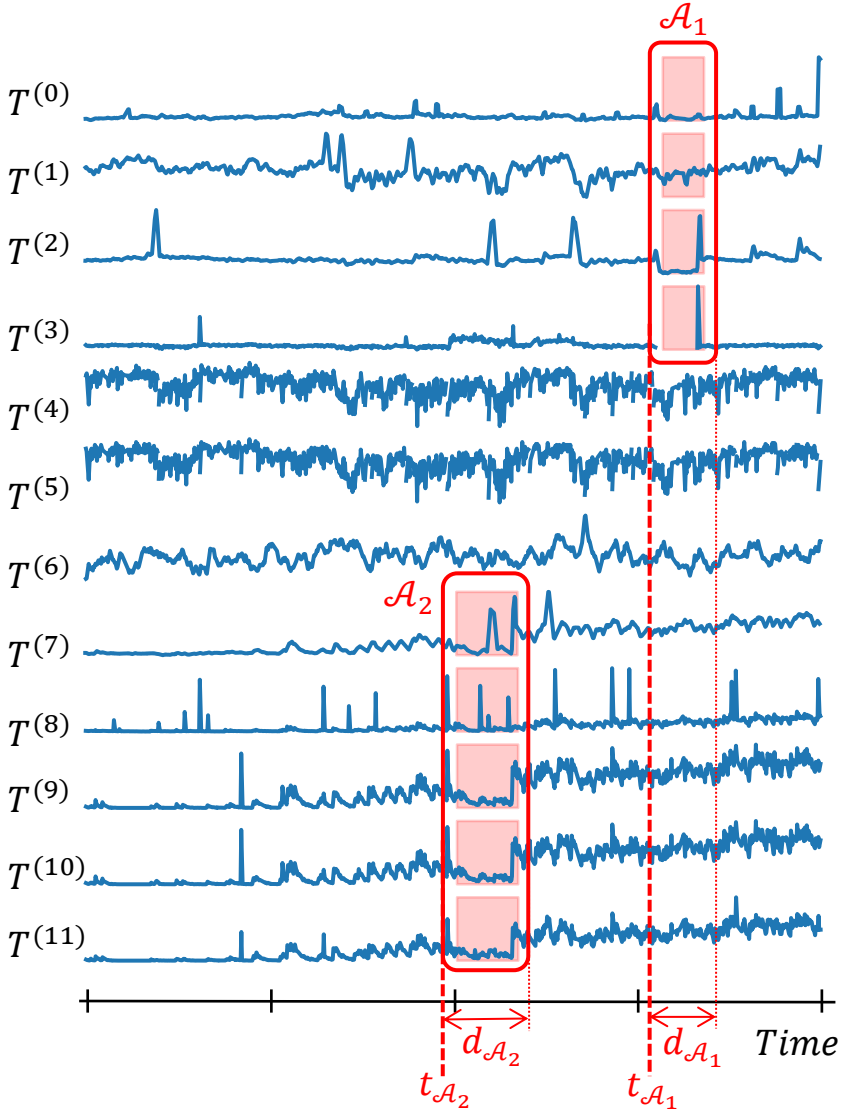
(a) Example of time series in **SHAD**

*12 dimensions among **171 in total***



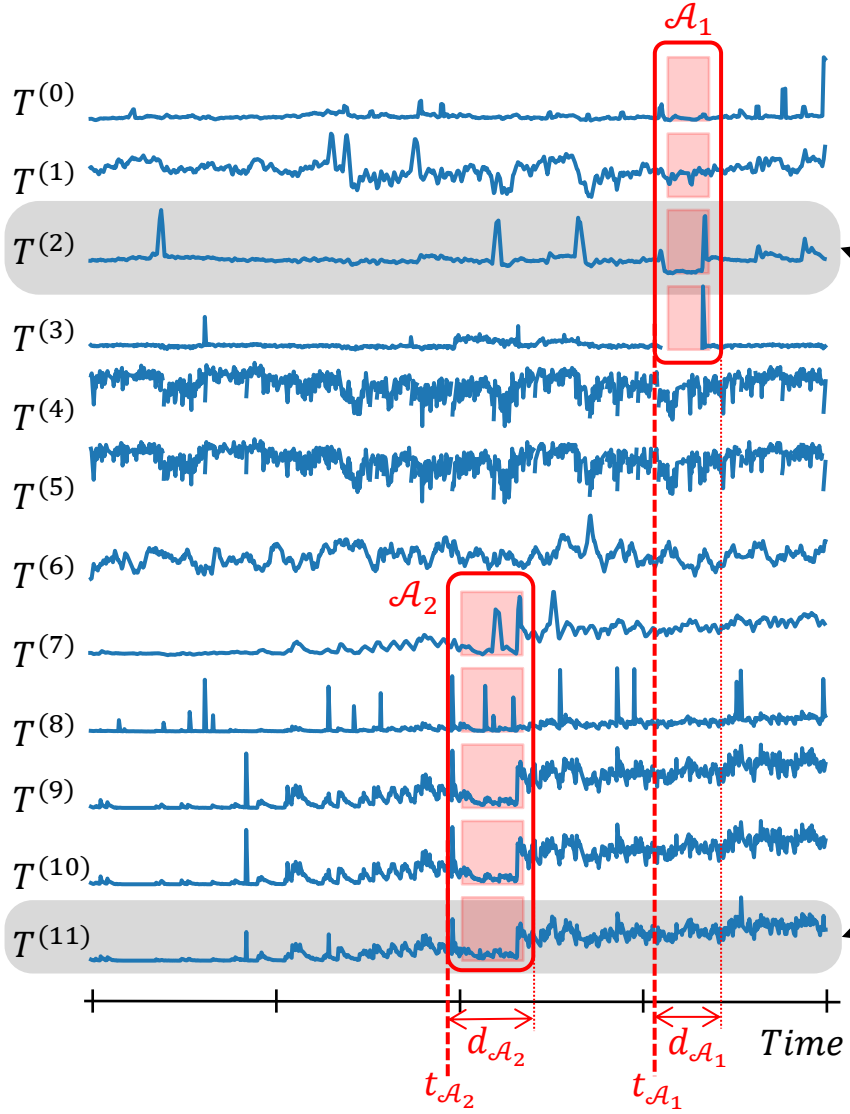
(a) Example of time series in **SHAD**

*12 dimensions among **171 in total***



(a) Example of time series in **SHAD**

*12 dimensions among **171 in total***



Anom. presence: *Yes*

Anom.Category: *Asynchronous*

Detection Label | Explanation Label | Textual Interpretation

Annotations for $T^{(2)}$

Metric Name	Metric Info
Store3-ssd02-disk_io_time_seconds	*SSD used for S3 metadata and small object storage (s/s) Fraction of time spent on I/O (1=saturated)*

Anomaly Info ($\mathcal{A}_1$)	Type	*Storage Node*	Sensors
2 Storage Node killed on Store 3 at time $t_{\mathcal{A}_1}$ = 14:00 for $d_{\mathcal{A}_1}$ =1 hour	Criticality	*High*	$T^{(0)}, T^{(1)}, T^{(2)}, T^{(3)}$
	Duration	*1.00 h*	
	Degradation	*None*	

⋮

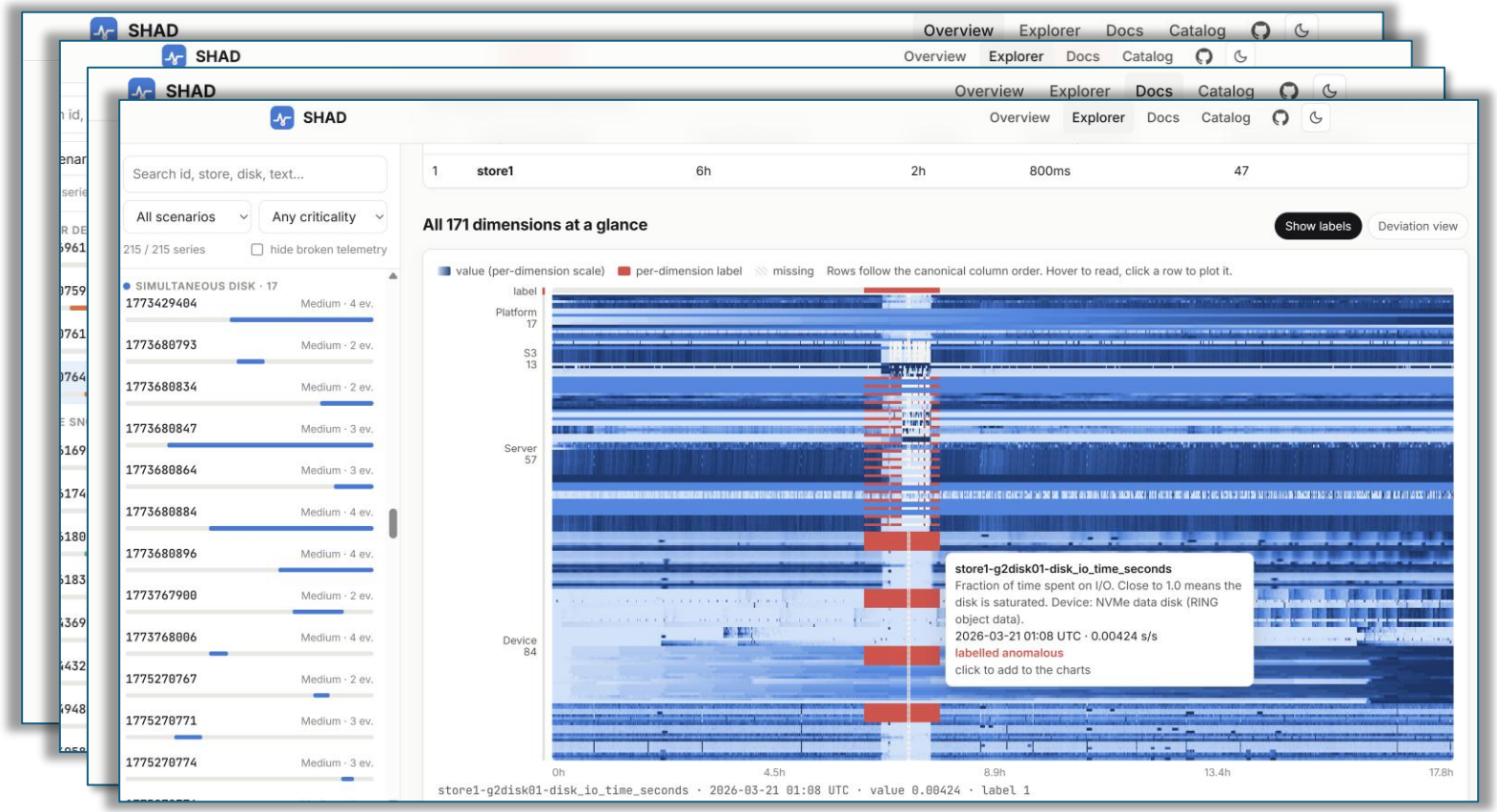
Annotations for $T^{(11)}$

Metric Name	Metric Info
Store1-g2disk01-disk_read_bytes	*NVMe data disk, where object data is stored (Bytes/s) Disk read throughput*

Anomaly Info ($\mathcal{A}_2$)	Type	*Storage Node*	Sensors
1 Storage Node killed on Store 1 at time $t_{\mathcal{A}_2}$ = 9:00 for $d_{\mathcal{A}_2}$ =1 hour 30 minutes	Criticality	*High*	$T^{(7)}, T^{(8)}, T^{(9)}, T^{(10)}, T^{(11)}$
	Duration	*1.30 h*	
	Degradation	*None*	

SHAD Explorer

- Time series in SHAD are explorable in the **online explorer**



Online Explorer

Experiments

RQ1: Detection

How do anomaly detection methods perform on SHAD, and how do they compare in terms of effectiveness?

RQ2: Explainability

Can TSAD methods achieve explainability through dimension-level attribution?

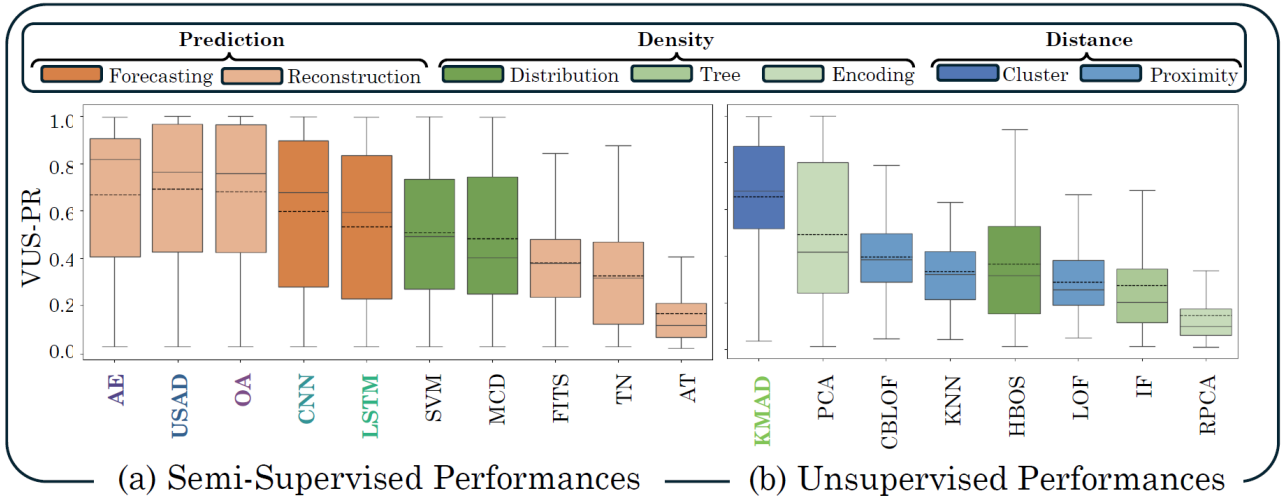
RQ3: Interpretability

Can frozen LLM models provide accurate anomaly interpretations?

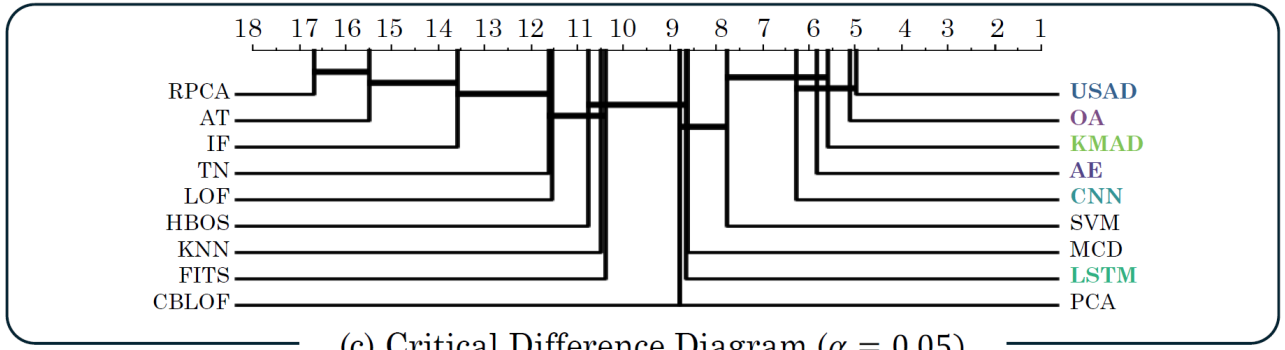
RQ1: Detection

- We assess the effectiveness of **18 state-of-the-art** detection **methods** from the TSB-AD benchmark [3]
- We include both **Semi-Supervised** (trained on nominal data) and **Unsupervised** approaches
  - We use the optimized hyperparameters from TSB-AD
- We **evaluate** performance using **VUS-PR** [4] **with a 25 point buffer**

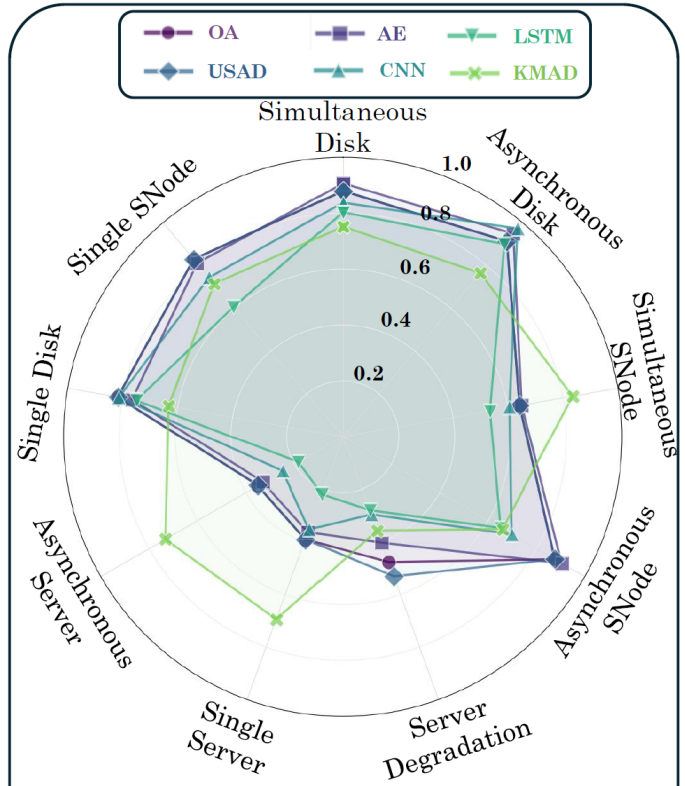
RQ1: Detection



(a) Semi-Supervised Performances — (b) Unsupervised Performances



(c) Critical Difference Diagram ($\alpha = 0.05$)



(d) Mean VUS-PR per Anomaly Type for Top 6 Methods

RQ1: Detect

SHAD remains non-trivial in Anomaly Detection, as no type of method manages to be exhaustively accurate on each anomaly type.

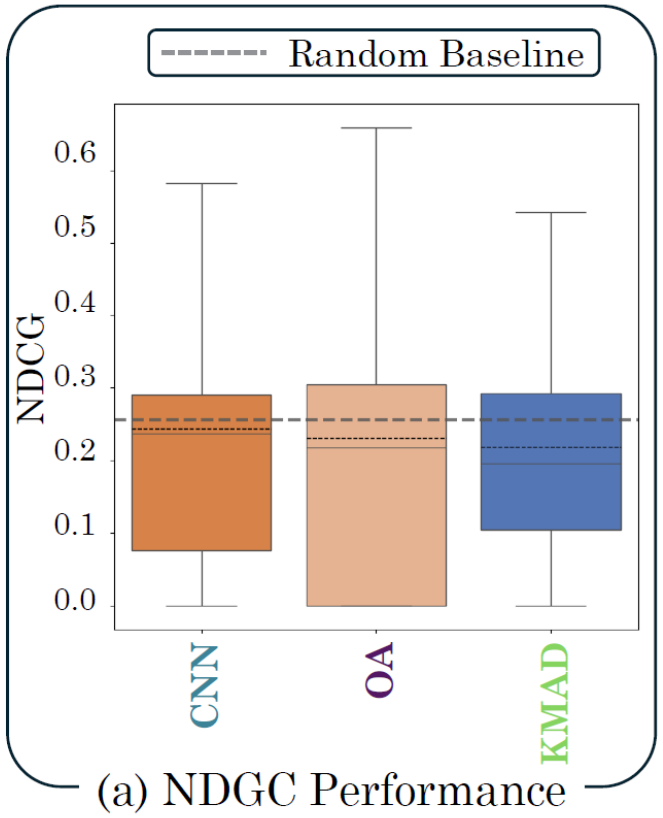
RQ2: Explainability

- We evaluate whether analyzing the **per-dimension contribution** in the anomaly score is sufficient to correctly identify anomalous dimensions.

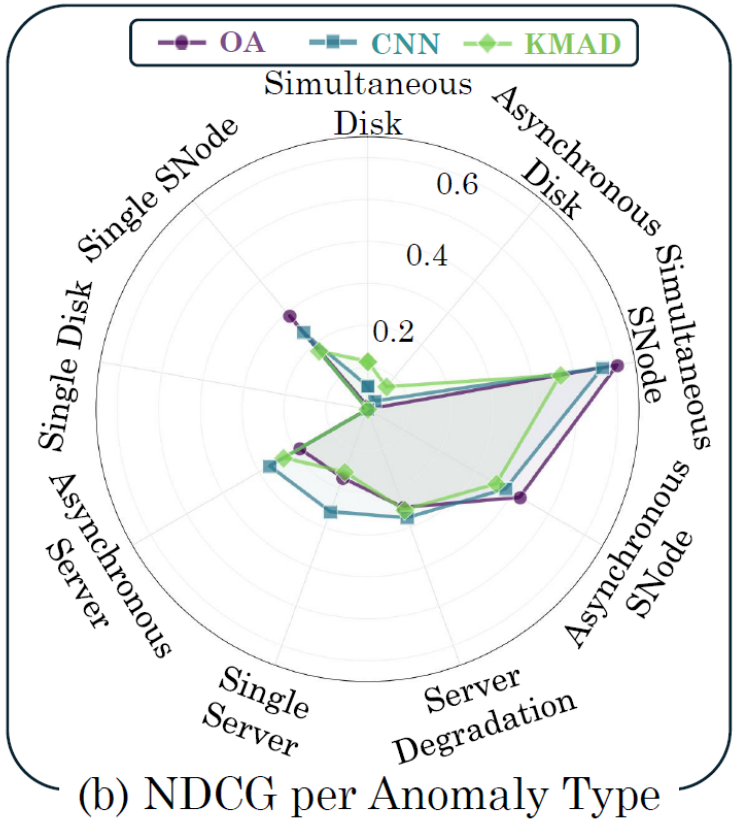
- We evaluate their performance using the **NDCG metric**, since this metric rewards well ranked features diagnosis.

- We select three representative methods and derive their feature-wise scores.
  - **OmniAnomaly** (Semi-Supervised, Reconstruction-based),
  - **CNN** (Semi-Supervised, Prediction-based)
  - **KMAD** (Unsupervised, Clustering-based)

RQ2: Explainability



(a) NDGC Performance



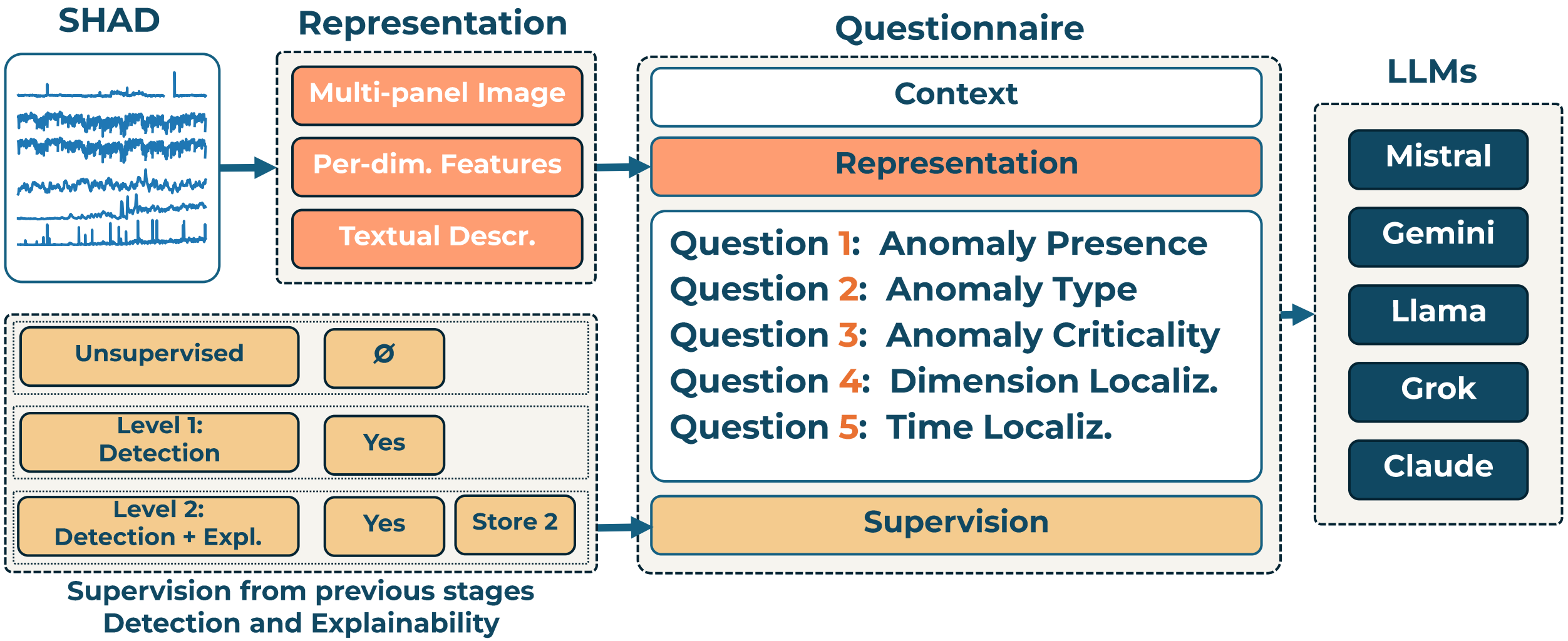
(b) NDCG per Anomaly Type

RQ2: Explain

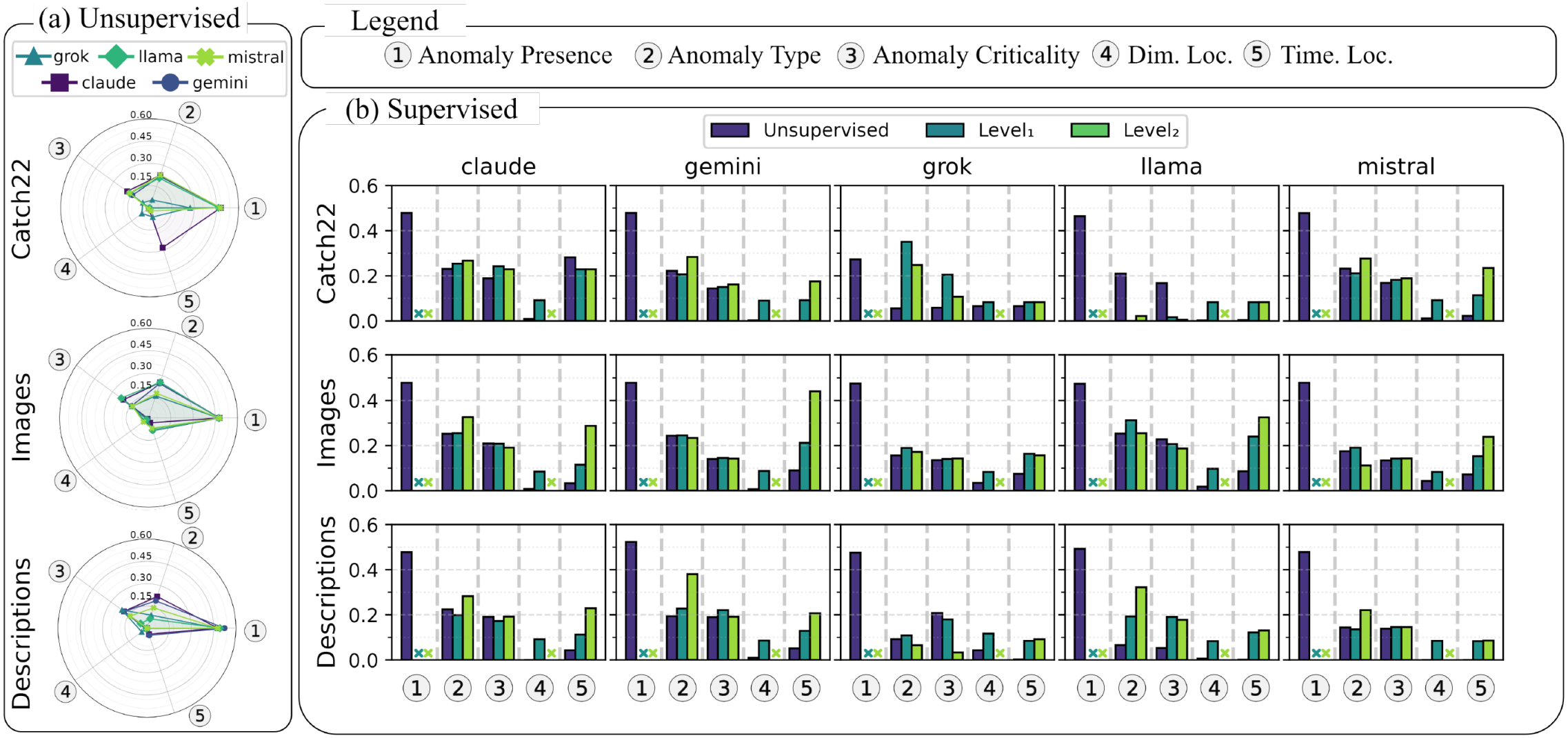
The inherent diversity and complexity of SHAD establish it as a non-trivial explainability testing ground that cannot be addressed with simple dimension contribution analysis.

RQ3: Interpretability

- We evaluate the ability of **frozen LLMs to interpret anomalies**.



RQ3: Interpretability



RQ3: Interpretability

Supervision	Task	Mean Gain
Level 1	Anomaly Criticality	0.0128
Level 2	Anomaly Criticality	-0.0078
Level 1	Dimension Localization	**0.0719**
Level 2	Dimension Localization	N.A. (trivial)
Level 1	Anomaly Presence	N.A. (trivial)
Level 2	Anomaly Presence	N.A. (trivial)
Level 1	Temporal Localization	**0.0793**
Level 2	Temporal Localization	**0.1446**
Level 1	Anomaly Type	0.0218
Level 2	Anomaly Type	0.0477

Having a **detector** before the interpretability step increases the **ability of identifying anomalous dimensions**

Having a **detector and an explainer** before the interpretability step increases temporal localization

RQ3: Interpret

Zero-shot, single-call LLMs cannot properly interpret SHAD's anomalies. Adding info from detection and explanation provides some benefits, but challenges still remain

Conclusion & Future Work

- We proposed **SHAD**, the **first TSAD benchmark** that **supports the evaluation of a complete TSAD pipeline**

- In addition, **we provided preliminary experiments** to evaluate Detection, Interpretability and Explainability

- **Future work includes:**

  - (**Detection**) Investigate if detectors can identify anomalies by looking at not directly affected dimensions
  - (**Explainability**) Design tailored explainability methods
  - (**Interpretability**) Understanding how to provide **outputs from the previous steps as context for LLMs**

THANK YOU!

Paper

GitHub Repo

Online Explorer

ENS

UNIVERSITÉ PARIS SCIENCES & LETTRES

Inria

SCALITY

[1] Boniol, P. et al. (2024). Dive into time-series anomaly detection: A decade review. arXiv:2412.20512.
[2] Xie, Z. et al. (2024). Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. arXiv:2412.03104.
[3] Liu, Q., & Paparrizos, J. (2024). The elephant in the room: Towards a reliable time-series anomaly detection benchmark. Advances in Neural Information Processing Systems, 37, 108231-108261.
[4] Paparrizos, J. et al. (2022). Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. VLDB.