# Paper Review: *A Picture is Worth A Thousand Numbers: Enabling LLMs Reason about Time Series via Visualization*

Review by Jason Gillette

## Introduction

This paper deals with LLM reasoning over time series data. LLMs historically struggle with time series tasks. The authors of this paper offer a novel approach that involves using data visualizations with a multi-modal model to reason on time series data. Specific contributions involving this approach include two artifacts, the prompting approach, called **VL-Time**, that includes visualized time series data and natural language guided reasoning, and an evaluation framework called **TimerBed** that improves on existing evaluations for reasoning on time series data.

## Research Questions

The authors approached this paper with (2) primary research questions.

1. Can LLMs be reused for time-series reasoning? If not, how can we enable them?
2. How to effectively evaluate LLMs on time series reasoning (TsR)?

## Contributions

The paper makes two main contributions.

1. **TimerBed**: A comprehensive evaluation suite stratified by reasoning patterns (simple deterministic, complex deterministic, and probabilistic). It provides real-world tasks, combinations of reasoning strategies (zero-shot, few-shot / in-context learning, chain-of-thought), and comparison anchors, including supervised models.
2. **VL-Time**: A prompt-based solution that integrates visualization with natural language guidance for task specific reasoning. This solution is a two-stage process.
    - planning (choosing visualization type and reasoning cues)
    - solving (executing TsR using multi-modal LLMs)—compresses information and highlights features that are otherwise obscured in raw numerical form.

## Methodology

The methodology is built on two components. First, the authors present an evaluation framework called **TimerBed** to compare their novel method to existing works in time series reasoning tasks. Next they present their novel methodology called **VL-Time**.

### TimerBed

The authors identify (3) main obstacles with existing evaluations which they seek to address.

1. Task Structure and Datasets

2. LLMs Reasoning Strategies, e.g. zero-shot, few-shot / in-context learning, chain-of-thought.
3. Comparison Anchors

**TimerBed** organizes TsR into three reasoning patterns, each mapped to real-world tasks:

1. Simple Deterministic Reasoning – one-to-one mapping between input and label
    - Right Whale Call detection
    - Transient Electromagnetic
    - Event classification
2. Complex Deterministic Reasoning – many-to-one mapping requiring integration of multiple features
    - ECG arrhythmia diagnosis
    - EMG signal classification
3. Probabilistic Reasoning – reasoning under uncertainty with unobserved variables
    - Human Activity Recognition
    - Computer Type Usage classification

TimerBed enables evaluation across the following reasoning strategies.

- zero-shot (ZST)
- chain-of-thought (CoT)
- few-shot in-context learning (ICL)

The authors included randomized guessing and diverse set of anchor models within TimerBed to ensure an accurate benchmark against existing works. Anchor models include (8) supervised models, (2) open-source models, e.g. *Qwen2.5*, and (2) closed-source models, e.g. *GPT-4o*.

## VL-Time

The study highlights two main causes of LLM failures on TsR that they address with **VL-Time**.

1. poor feature extraction from numerical inputs.
2. excessive tokenization overhead in long sequences of time series data. Visualization provides a solution by compressing inputs and surfacing domain-relevant features.

**VL-Time** was inspired by the human approach to reason on time-series data via visualization. Intuition on why visual reasoning can outperform numerical feature extraction is as follows.

- Pictorial features are more immediately discernable.
- Direct representation of time series by placing on an axis.
- Visual alignment of co-occurring features with clear labels or colors depicting cross-domain.
- Proper representation of frequency.

Visualization further enables time series reasoning by effectively compressing time series data. Numerical representation may produce hundreds to thousands of tokens, where a visualization will compress the same data into an image; the authors illustrate one example where 60k tokens of numerical representation are compressed into an 85 token image.

VL-Time executes in (2) stages.

1. Planning Stage
    - Prompt LLM as domain expert

    - Choose between time-based or frequency-based visualization
    - Propose reasoning clues
2. Solving Stage - Prompt the multi-modal LLM to perform reasoning task using visualization image in-context.

# Results & Analysis

## Anchor Results

How effective are LLMs for TsR under different reasoning patterns and strategies?"

- Zero-shot prompting results are near the random guessing baseline.
- Chain-of-thought (CoT) consistently outperforms other methods.
- Few-shot / in-context learning (ICL) suffers from performance degradation.
- Uni-model LLMs out perform Multi-modal LLMs suggesting language degradation in multi-modal models.

Primary cause of failures is poor feature extraction as observed when models are prompted to return their feature extraction plans. Types of feature extract relative to unique domain tasks are identified in (4) categories.

1. Pictorial Features - Patterns within visualization that are difficult to capture numerically.
2. Time-aware Features - Timestamp features that are lost in the model's data representation.
3. Cross-dimensional Features - Simultaneous changes across features that lose their temporal alignment as model inputs.
4. Frequency-domain Features

The next cause of failure is excessive context length leading to performance degradation. Hundreds of timestamps in a given dataset is often as 12x in tokens. This further degrades ICL prompting methods as there is limited context length available for examples, e.g. an example may contain 60k tokens while the model is limited to a 120k context length.

## VL-Time Performance Results

Results were determined using the OpenAI GPT-4o model. The significant results observed include the following.

- VL-Time enabled the use of zero-shot methods with multi-modal LLMs by demonstrating improvements; surpassing random guessing on all tasks and out performing supervised anchor models on four out of six tasks.
- Enables few-shot in-context learning methods by compressing in-context samples; outperforming supervised models on all tasks with simple and complex deterministic patterns while matching supervised models on all probabilistic reasoning tasks. However, VL-Time with its multi-modal model enables probabilistic reasoning with significantly less tokens. IN other words, it matches performance with less cost that numerical representation.

# Discussion

The VL-Time method demonstrates significant improvements over existing time series reasoning methods. The most promising aspect briefly mentioned by the authors is the potential cost savings due

to the data compression offered by the visualization. The benefits have potential for applicability beyond the presented use cases and may be promising future research directions. However, the method in practice as depicted by the others puts the time consuming burden of data visualization on the end user. It is unclear if the authors aim automate this task through an additional step in VL-Time. They conduct ablation studies to identify key components of the visualization that contribute to increased performance, so the parameters for data visualization are present for extension.

### Questions

1. What is the accepted context length of each model evaluated, and how many examples were included in few-shot cases?
2. Could VL-Time be extended into a pipeline where the LLM itself generates the required visualization from numerical input (e.g., using Python code execution)?
3. How does visualization generalize beyond classification and would it support anomaly detection tasks equally well?
4. Could the VL-Time methodology be applicable to other reasoning tasks such as static classification via clustering visualizations?

## Reference

A Picture is Worth A Thousand Numbers: Enabling LLMs Reason about Time Series via Visualization (Liu et al., NAACL 2025)