



# VETime: Vision Enhanced Zero-Shot Time Series Anomaly Detection

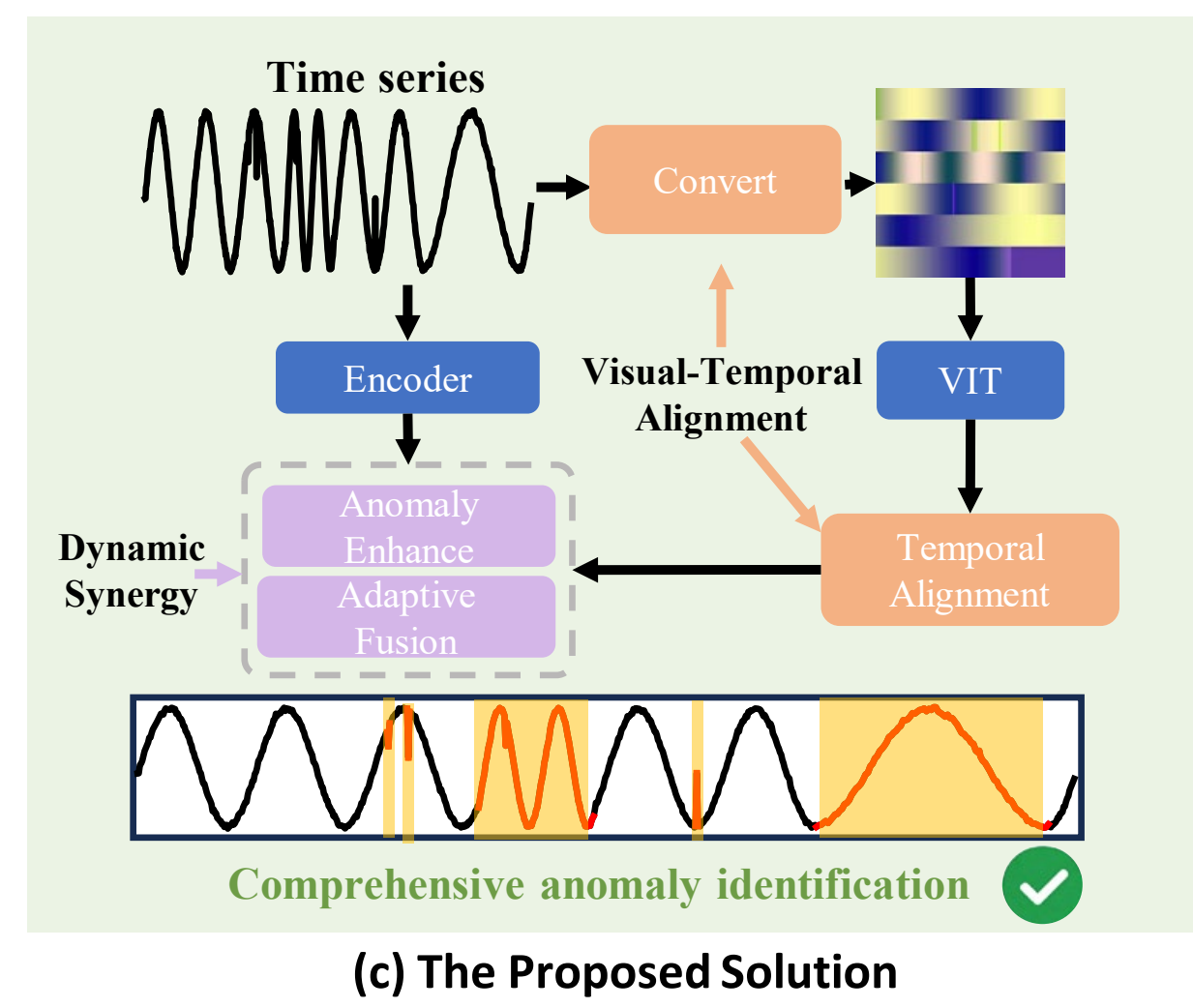
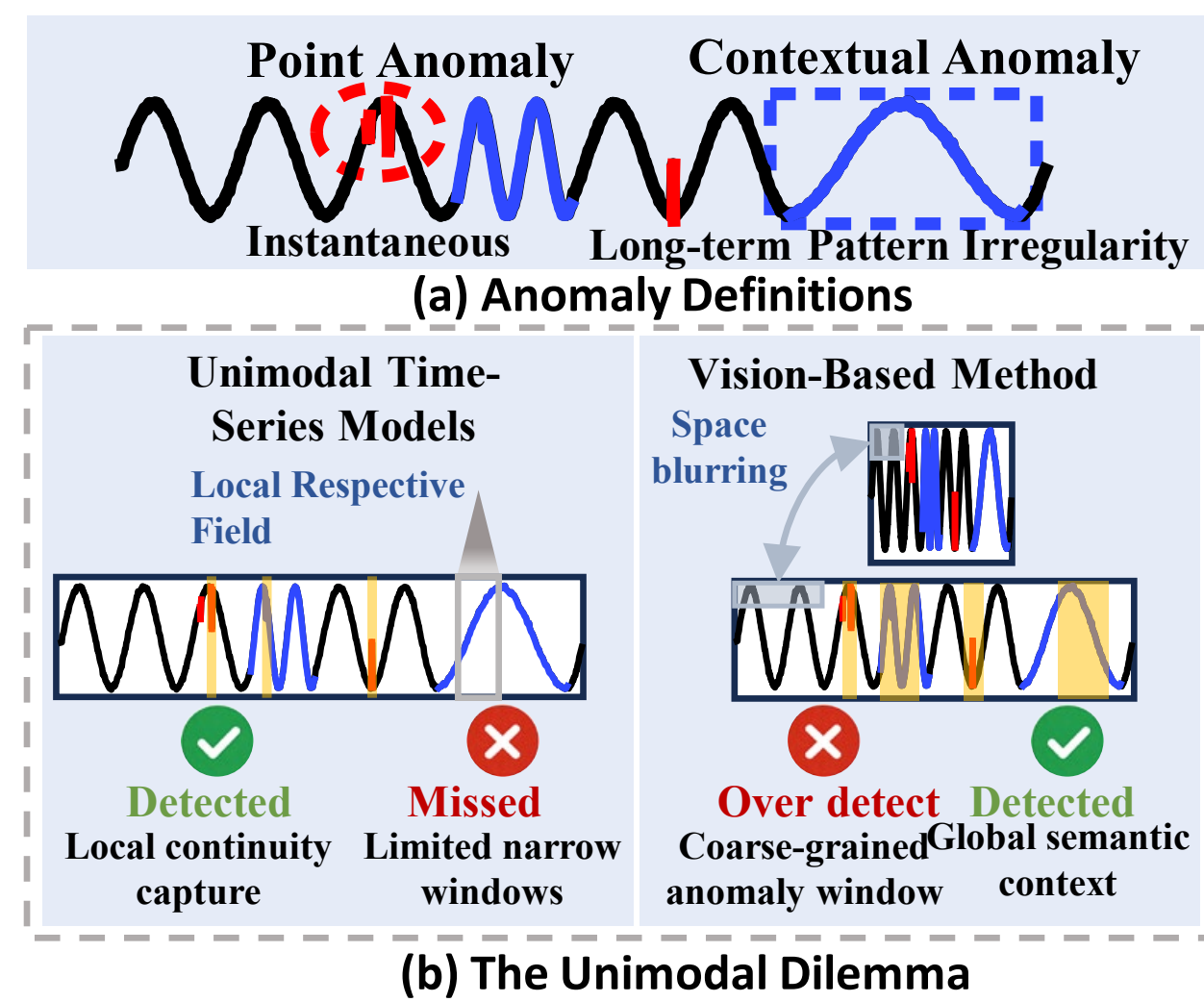
Yingyuan Yang<sup>1</sup> Tian Lan<sup>1</sup> Yifei Gao<sup>1</sup> Yimeng Lu<sup>1</sup> Xuming An<sup>1</sup>  
Wenjun He<sup>2</sup> Meng Wang<sup>2</sup> Chenghao Liu<sup>3†</sup> Chen Zhang<sup>1†</sup>

<sup>1</sup>Tsinghua University <sup>2</sup>Huawei 2012 Lab <sup>3</sup>Datadog AI Research <sup>†</sup>Corresponding authors



## Problem Definition and Contribution

**Task.** Zero-shot time-series anomaly detection must identify rare deviations and precisely localize their timestamps in unseen target domains, without target-specific training, tuning, or labels.



### Two Anomaly Regimes

- **Point anomalies:** abrupt, short-lived; require fine-grained local sensitivity.
- **Contextual anomalies:** violate long-range structure; require global contextual view.

### The Competency Gap

**Temporal models** preserve timestamp detail but remain locally constrained. **Visual models** reveal global patterns but lose exact timestamp alignment.

### Key Contributions

#### 1. Unified Framework

First zero-shot TSAD integrating visual and temporal features for dense anomaly localization.

#### 2. Traceable Shared Timeline

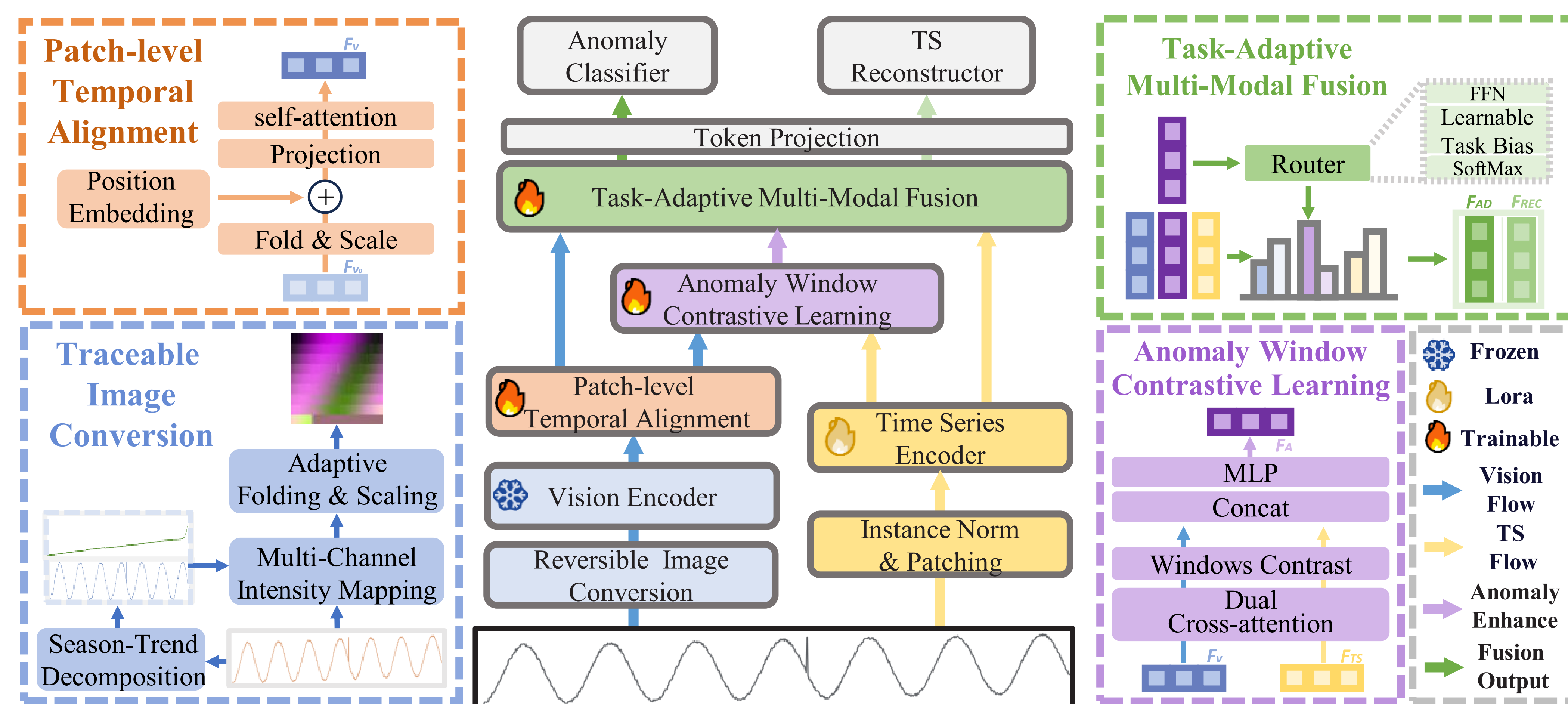
TIC and PTA retain rich visual context while restoring precise timestamp-to-patch correspondence.

#### 3. Adaptive Multimodal Synergy

AWCL and TMF learn complementary cues and dynamically route experts per patch and task.

## Method

**Core idea.** Convert while preserving temporal structure, align visual patches back to the shared timeline, learn local/global anomaly cues, and route complementary features for each patch and task.



#### 1. TIC

Raw / trend / remainder  $\rightarrow$  RGB; adaptive folding exposes cross-period structure.

#### 2. PTA

Map visual patches back to the ordered timeline shared with temporal tokens.

#### 3. AWCL

Intra-window contrast sharpens points; inter-window contrast captures context.

#### 4. TMF

Route temporal, visual, and anomaly-enhanced experts by patch and task.

## Experiments & Results

### Strict zero-shot protocol

#### 0-shot

no target-domain tuning

#### 0.5B

synthetic training points

#### 26

unseen public datasets

#### 4

complementary metrics

### Zero-shot baseline comparison

**Zero-shot TSAD: Univariate**  
24/44 first-place Avg. rank 2.05

**Zero-shot TSAD: Multivariate**  
34/60 first-place Avg. rank 1.58

### Vision-based baselines

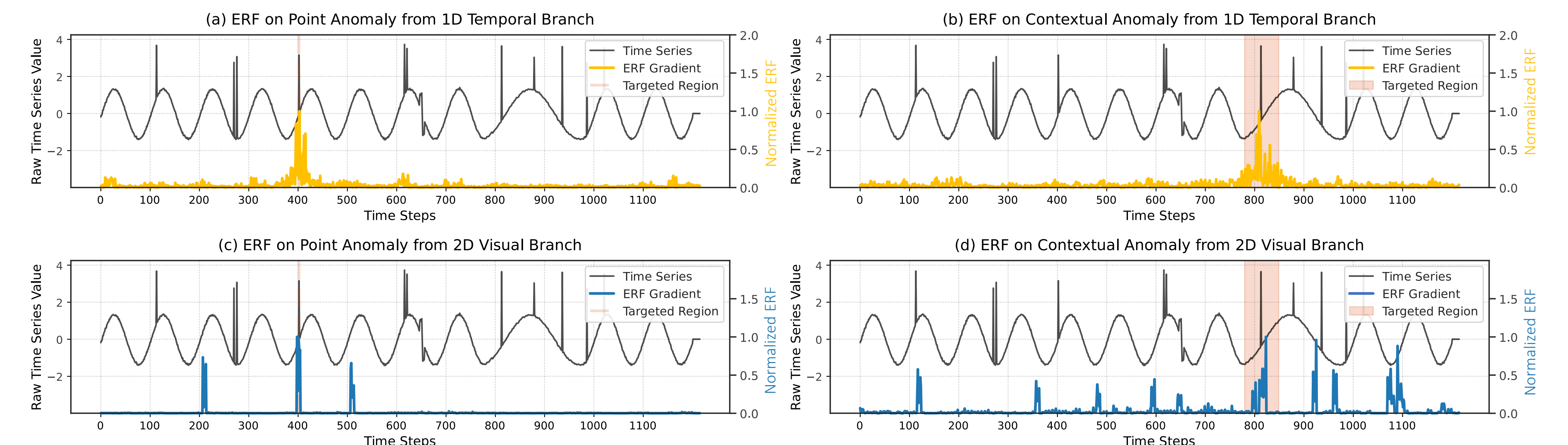
Best across 4 datasets  $\geq 43\times$  faster

### Performance across anomaly types

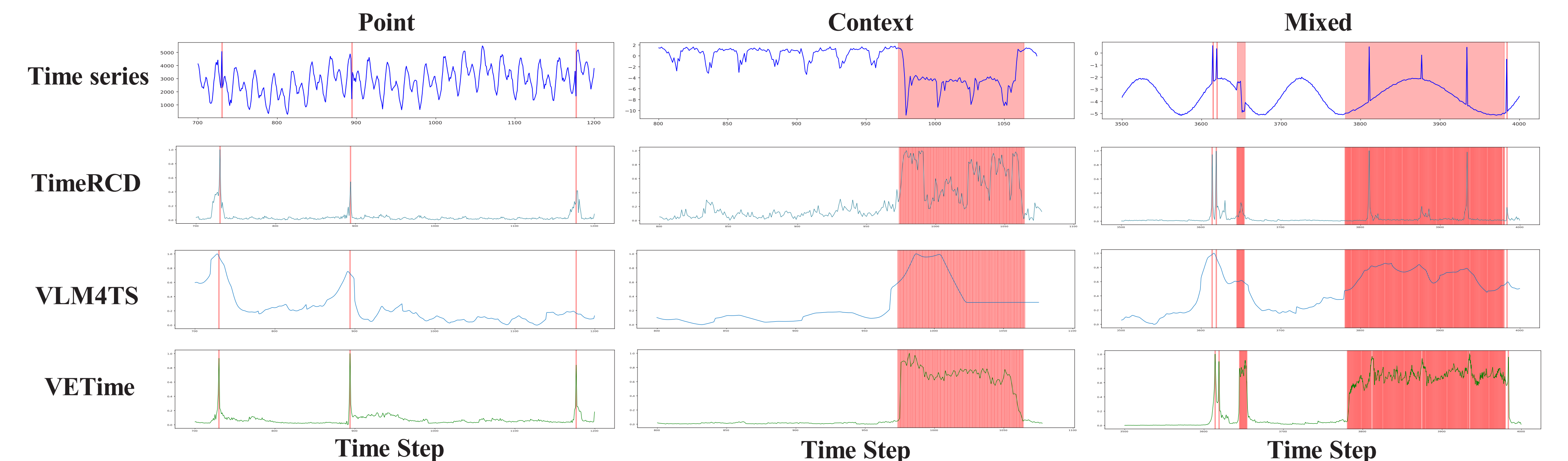
| Model   | Point         | Context       | Mixed         |
|---------|---------------|---------------|---------------|
| VETime  | 93.20 / 59.33 | 91.79 / 57.59 | 87.78 / 44.61 |
| TimeRCD | 74.84 / 61.85 | 73.84 / 44.93 | 73.03 / 42.90 |
| VLM4TS  | 67.11 / 30.57 | 61.94 / 32.10 | 60.79 / 25.61 |

Aff-F1 / VUS-PR (%)

### Why Multimodality Helps?



### Precise localization



### Ordered temporal-to-visual mapping

