SE-STGAE: A Shapelet-Enhanced Spatiotemporal Graph Autoencoder for Early Anomaly Detection

En Wang[1,2], Fujia Zhou[1,2], Hao Du[1,2], Wenbin Liu[1,2], Jie Wu[3,4]

[1] College of Computer Science and Technology, Jilin University, China

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China

[3] China Telecom Cloud Computing Research Institute, China

[4] Department of Computer and Information Sciences, Temple University, USA

Email: wangen@jlu.edu.cn,{zhoufj24,duhao22}@mails.jlu.edu.cn, liuwenbin@jlu.edu.cn, jiewu@temple.edu

***Abstract*—Early anomaly detection in networked sensing is crucial for maintaining the Quality of Service (QoS), especially under strict False Positive Rate (FPR) budgets where alarms must be early and reliable. Existing methods typically learn normal spatiotemporal patterns from multi-node sensing streams and raise alarms using reconstruction deviations in a one-class manner. However, in real freeway monitoring, anomalies exhibit complex spatiotemporal propagation and the early anomaly signals are sparse. Meanwhile, traffic streams contain diverse normal fluctuations that may look abnormal, making them hard to distinguish from real incidents. Together, these factors make anomaly propagation hard to capture under strict low-FPR early-warning settings. To address these challenges, we propose SE-STGAE, a shapelet-enhanced spatiotemporal graph autoencoder. It first uses shapelets to extract localized segment-level precursors, then integrates temporal precursors with propagation-aware spatial dependencies to improve robustness to normal fluctuations, and finally applies a tail-emphasized Top-K scoring rule with pure-normal calibration for reliable early alarming. Extensive experiments on real-world freeway datasets show consistent gains over state-of-the-art baselines across datasets and FPR budgets, delivering earlier detection with localized evidence.**

***Index Terms*—Freeway incidents, QoS-aware detection, spatiotemporal data, shapelet learning.**

I. INTRODUCTION

In networked sensing and control systems, early anomaly detection is crucial for improving the Quality of Service (QoS), since abnormal events can disrupt system operation [1]. It must provide early warnings under a predefined False Positive Rate (FPR) to suppress the impact of anomaly propagation [2], thereby ensuring the stability of sensing services. A representative application is large-scale freeway monitoring, where incidents constitute critical anomalies that may trigger congestion propagation and service degradation [3]. Therefore, for such time-sensitive sensing services, reliable early warning under a low FPR is essential for QoS improvement [4].

Building on high-resolution roadside sensing streams, recent studies have explored data-driven early anomaly detection [5]. A dominant paradigm employs spatiotemporal graph [6] autoencoders and raises alarms using node-level reconstruction deviations. However, these methods often overlook an important fact: early anomalies do not emerge as isolated point-

979-8-3195-1703-6/26/$31.00 © 2026 IEEE

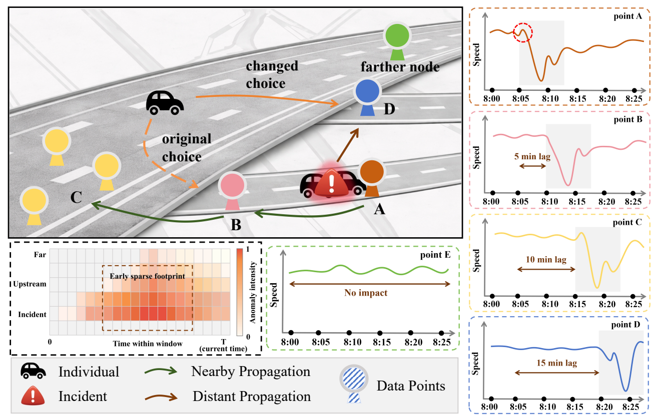


Fig. 1: An intuitive example of an anomalous event and its spatiotemporal propagation.

wise signals, but propagate across both time and space. Node-level modeling is insufficient to capture anomalous dependencies among nodes [7], calling for segment-level modeling to capture early anomaly emergence and propagation. As shown in Fig. 1, when an incident occurs at point A on the ramp, nearby segments first exhibit a sharp speed drop, which quickly triggers queuing and congestion at the upstream point B. As time progresses, the anomaly further spreads and affects point C on the adjacent mainline, causing traffic speed to decrease, while farther nodes remain near normal within the same time window. This observation indicates that early anomalies emerge as localized, short-lived patterns and propagate across time and space, which is important for timely warning and reliable alarms [8].

To develop a method that can model such localized-and-propagating precursors, three unique challenges arise:

1) How to capture complex anomaly propagation under sparse early precursors? In real traffic networks, spatiotemporal anomaly propagation is complex and hard to characterize. Fig. 1 shows that when congestion occurs at ramp A, individuals who originally planned to exit via this ramp may choose to leave at the next exit to avoid the congestion. Such changes mean that the impact of the anomaly at point A not only spreads through nearby areas, but can also propagate beyond physical adjacency to the distant point D [9], slowing down traffic speed in its surrounding road segments. This indicates that anomaly propagation is driven not only by spatial proximity [10], but also by topology-agnostic factors, which is difficult to capture using only a static adjacency structure [11]. Meanwhile, early anomaly signals are localized and sparse, further increasing the difficulty of identifying propagation chains and key precursors.

2) How to distinguish real incidents from diverse normal fluctuations in traffic networks? In real-world scenarios, spatiotemporal data naturally exhibit significant variability. For example, during holidays, traffic demand increases and queues at some exits cause a sustained speed drop, which further leads to congestion between ramps and the mainline. Such common spatiotemporal fluctuations may look abnormal but are not incidents. More importantly, since normal fluctuations dominate and incidents are rare, early detection can confuse weak incident precursors with normal variations, causing false alarms or missed detections. Therefore, reliably separating normal fluctuations from true anomalies under complex spatiotemporal variability remains a key challenge.

3) How to identify and prioritize the most informative precursors from all signals? While early anomalous precursors can be extracted at the segment-level, only a few precursor signals are truly informative for early warning. In the early stage of an incident, only a few locations around the incident site show clear deviations; similarly, within a detection window the most discriminative signals often appear toward the time of official confirmation. Uniform aggregation can therefore dilute these critical signals with many weak or irrelevant ones, weakening anomaly features and delaying alarms. This calls for a mechanism that highlights the most informative precursor patterns to support timely and reliable alarming.

To address these challenges, we propose Shapelet-Enhanced SpatioTemporal Graph AutoEncoder (SE-STGAE) for early freeway anomaly detection under strict FPR budgets. **1) To capture early, localized precursors under complex spatial propagation**, we introduce Shapelet-Based Local Pattern Encoding to extract local shape deviations as precursor evidence, and integrate Dynamic Multi-Hop Spatial Enhancement with dynamic adjacency learning and multi-hop propagation, allowing spatial dependencies to adapt over time and across road segments. **2) To distinguish real incidents from normal fluctuations**, we develop a Cross-View Attention and Residual Field Fusion module that aligns temporal local patterns with spatial propagation patterns and fuses them through a time-conditioned residual field, improving robustness to non-stationary regimes and regular congestion variations. **3) To identify and prioritize the most informative precursors**, we design Tail-Emphasized Top-K Scoring for both training and inference to prioritize the most informative deviations while keeping false alarms controllable. Overall, SE-STGAE preserves the reconstruction-based one-class paradigm while providing clear, localized evidence for early anomaly warning.

Our work has the following contributions:

- We address QoS-aware early anomaly detection under strict FPR budgets in networked freeway sensing, and propose SE-STGAE, a shapelet-enhanced spatiotemporal graph autoencoder.
- We design a spatially-aware shapelet-based temporal encoder, where shapelets act as localized temporal prototypes and are explicitly embedded into a spatiotemporal modeling framework, enabling robust precursor extraction under noisy local fluctuations.
- We propose an adaptive spatiotemporal pipeline with dynamic multi-hop spatial enhancement and cross-view attention with residual field fusion, and develop an FPR-oriented alarming rule via tail-emphasized Top-K scoring with pure-normal calibration.
- We evaluate the proposed method on three real-world freeway datasets and show improvements over baselines under low-FPR early-warning protocols, demonstrating the effectiveness and robustness of SE-STGAE.

II. Related Work

A. Reconstruction-based Anomaly Detection

Reconstruction-based one-class detectors are widely used for freeway traffic anomaly detection, as they learn normal spatiotemporal dynamics from pure-normal data and flag anomalies via elevated reconstruction errors. Recent benchmarks such as FT-AED [12] further highlight lane-level, high-resolution sensing streams and early-event evaluation settings. To model spatial dependencies, many studies formulate traffic networks as spatiotemporal graphs [13] and adopt graph autoencoders or generative frameworks [14, 15] to capture normal patterns; representative lines also explore adversarial and cycle-consistent generators for lane-level detection. However, much of the existing literature focuses on improving overall reconstruction fidelity or reporting threshold-agnostic performance, whereas operational deployment often requires calibrated decisions at strict low-FPR operating points under QoS budgets [16]. In addition, early-stage deviations are typically sparse and localized, and global aggregation of deviations may mask weak precursors, yielding limited localized evidence to support verification and decision auditing [17].

B. Shapelet Learning for Localized Patterns

Shapelets are subsequences that capture salient local patterns in time series and provide explicit localization of signals [18]. Beyond supervised classification, recent studies investigate shapelet-based unsupervised representation learning for multivariate time series [19], aiming to extract general-purpose local features without labels [20]. Shapelet learning has also been extended with network-regularized or adaptive adjacency constraints to exploit inter-variable relations in multivariate signals [21], such as incorporating graph-based relationships into adversarial shapelet learning [22]. Nevertheless, shapelet learning remains underexplored in spatiotemporal graph anomaly detection, where precursor patterns must be aligned across nodes and remain coherent under spatial propagation and non-stationary dynamics. These observations motivate integrating shapelet-based localized signals into reconstruction-based spatiotemporal graph detectors to improve sensitivity to early precursors, while supporting clear and localized decision cues under strict FPR constraints [23].

III. System model and Problem formulation

A. System Model

We consider a freeway sensing system that provides QoS-aware early anomaly warning under strict FPR constraints. The system consists of N fixed roadside stations. Each station reports a C-dimensional measurement vector at a fixed sampling interval. We model the sensing network as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and denote its topology by a static adjacency matrix:

$$\mathbf{A}_0 \in \mathbb{R}^{N\times N}, \tag{1}$$

where $|\mathcal{V}| = N$ and $\mathbf{A}_0$ is derived from road topology or spatial proximity (connectivity or distance along the freeway).

Spatiotemporal Data Stream. Let $\mathbf{x}_t \in \mathbb{R}^{N\times C}$ denote the C-dimensional observations collected from all N stations at time index t. We operate on sliding windows of length T:

$$\mathbf{X}_t = \{\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t\} \in \mathbb{R}^{N\times T\times C}, \tag{2}$$

where T is the window length. We further provide a time-context feature $\mathbf{e}_t$ (time-of-day encoding) as an auxiliary conditioning signal.

Event Annotations and One-Class Protocol. Let $y_t \in \{0, 1\}$ indicate whether time t is associated with an anomalous event. Importantly, y_t is not used for supervised learning; it is used only for filtering the training data and for evaluation. Since anomalous events are sparse and their impacts can propagate, we adopt a one-class protocol: the training set $\mathcal{D}_{\mathrm{tr}}$ contains only windows verified to be normal, ensuring a pure-normal learning setting. We also construct a normal validation set $\mathcal{D}_{\mathrm{val}}$ that follows the same protocol for threshold calibration.

B. Problem Formulation

Given a sensing window $\mathbf{X}_t$, the base graph $\mathbf{A}_0$, and the time context $\mathbf{e}_t$, our goal is to learn a reconstruction-based one-class detector that models normal spatiotemporal dynamics and produces QoS-compliant early-warning decisions under a target FPR budget. We parameterize the detector as an encoder–decoder model:

$$\mathbf{z}_t = f_\theta(\mathbf{X}_t; \mathbf{A}_0, \mathbf{e}_t), \qquad \widehat{\mathbf{X}}_t = g_\phi(\mathbf{z}_t, \mathbf{e}_t), \tag{3}$$

where $f_\theta(\cdot)$ maps the input window to a latent representation $\mathbf{z}_t$, and $g_\phi(\cdot)$ reconstructs the window $\widehat{\mathbf{X}}_t$.

One-Class Reconstruction Learning. We learn (θ, ϕ) on the pure-normal training set $\mathcal{D}_{\mathrm{tr}}$ by minimizing a reconstruction loss:

$$\min_{\theta,\phi} \; \mathbb{E}_{\mathbf{X}_t\sim\mathcal{D}_{\mathrm{tr}}} \left[\mathcal{L}_{rec}(\mathbf{X}_t, \widehat{\mathbf{X}}_t)\right], \tag{4}$$

where $\mathcal{L}_{rec}(\cdot)$ measures reconstruction loss and will be instantiated in Section IV to support QoS-aware early warning.

QoS-Aware Inference via FPR Calibration. At test time, we compute a window-level anomaly score:

$$s(\mathbf{X}_t) = \mathcal{S}(\mathbf{X}_t, \widehat{\mathbf{X}}_t), \tag{5}$$

where $\mathcal{S}(\cdot)$ quantifies how much the window $\mathbf{X}_t$ deviates from the reconstructed normal pattern $\widehat{\mathbf{X}}_t$. An alarm is raised if $s(\mathbf{X}_t) > \tau_\alpha$.

To meet a target FPR budget α, we calibrate the threshold τ_α using the normal validation set $\mathcal{D}_{\mathrm{val}}$ by setting τ_α to the (1 - α)-quantile of normal scores, so that the empirical FPR on normal data matches the desired operating point by construction. The specific design of $\mathcal{L}_{rec}(\cdot)$ and $\mathcal{S}(\cdot)$ will be detailed in Section IV.

IV. Method

A. Method Overview

As illustrated in Fig. 2, our framework consists of four components: Shapelet-Based Local Pattern Encoding, Dynamic Multi-Hop Spatial Enhancement, Cross-View Alignment and Residual Field Fusion, and Reconstruction and QoS-Calibrated Decision. Given an input window, we encode localized temporal patterns via shapelet matching with a lightweight one-hop refinement. Then, we learn a dynamic adjacency and perform multi-hop propagation to enhance spatial features. Following this, we align and fuse spatiotemporal representations via cross-view interaction and residual neural fusion to obtain a unified latent representation. A decoder reconstructs the input, and anomalies are scored by tail-emphasized Top-K reconstruction errors with an FPR-calibrated threshold. We summarize the workflow in Algorithm 1.

B. Encoder

1) Encoder Overview: The encoder maps each input window to a unified latent representation that captures localized precursors and spatial propagation. It consists of three stages: (i) Temporal–Spatial Pattern Representation extracts shapelet-based local patterns and applies a lightweight one-hop consistency refinement. (ii) Adaptive Spatial Enhancement learns an adaptive adjacency and performs multi-hop message passing to capture longer-range dependencies. (iii) Spatiotemporal Alignment and Fusion aligns the temporal view and the spatial view via cross interactions, further refines the fused features using a time-conditioned residual neural field to adapt to regime shifts. The fused representation is then passed to the decoder for reconstruction and online decision.

2) Shapelet-Based Local Pattern Encoding: This module extracts localized temporal patterns from each node's multivariate series, and stabilizes them with a lightweight one-hop spatial consistency refinement. Since anomalous precursors are often short and shape-specific and can be diluted by global aggregation, we adopt shapelet matching to explicitly preserve local pattern evidence in the representation.

Given a window $\mathbf{X}_t \in \mathbb{R}^{N\times T\times C}$, we denote the multivariate series of node i as $\mathbf{X}_{t,i} \in \mathbb{R}^{T\times C}$. For a shapelet length L, we generate local subsequences by sliding a length-L window with stride 1:

$$\mathbf{X}_{t,i}^{(u)} \in \mathbb{R}^{L\times C}, \quad u = 1, \ldots, W, \quad W = T - L + 1, \tag{6}$$

where $\mathbf{X}_{t,i}^{(u)}$ is the u-th local segment of node i. For each L, we learn a shapelet pool $\{\mathbf{S}_p^{(L)}\}_{p=1}^{P_L}$ with $\mathbf{S}_p^{(L)} \in \mathbb{R}^{L\times C}$. To

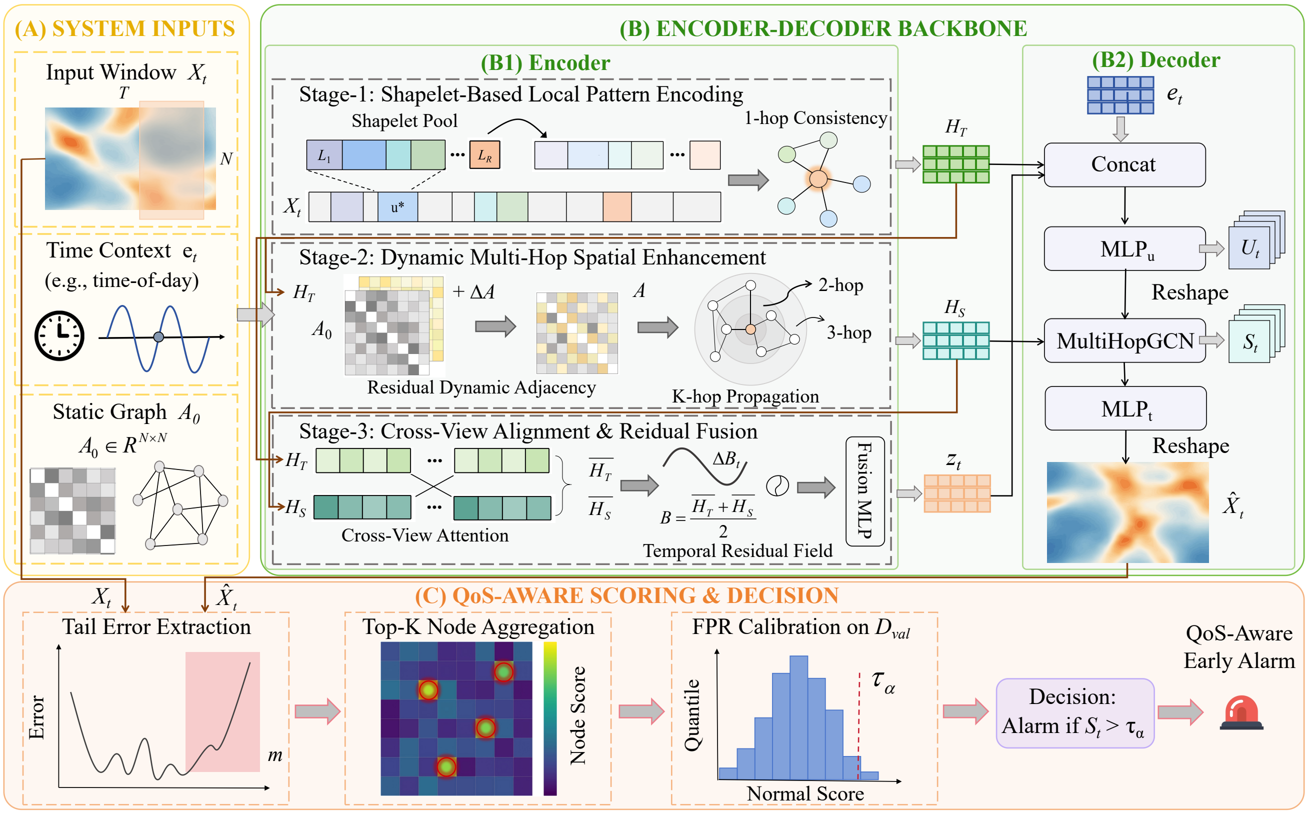


Fig. 2: The overview architecture of SE-STGAE, which consists of three components: (A) system inputs; (B) encoder–decoder backbone; (C) QoS-aware scoring & decision

find the best-matching subsequence of shapelet p in node i, we compute the distance (e.g., Euclidean distance after flattening channels):

$$d_{i,p}^{(u)} = \left\| \mathrm{vec}(\mathbf{X}_{t,i}^{(u)}) - \mathrm{vec}(\mathbf{S}_p^{(L)}) \right\|_2, \tag{7}$$

where $\mathrm{vec}(\cdot)$ concatenates all channels within the segment. We then summarize the window by the minimum distance for each shapelet:

$$h_{i,p}^{(L)} = \min_{u\in\{1,\ldots,W\}} d_{i,p}^{(u)}, \qquad u_{i,p}^{*(L)} = \arg\min_u d_{i,p}^{(u)}, \tag{8}$$

where $h_{i,p}^{(L)}$ is the shapelet response feature of node i, and $u_{i,p}^{*(L)}$ records the location of the best-fit subsequence for localization and interpretation. Intuitively, the $\min$ operator answers whether a prototype-like pattern appears anywhere in the window, while u^* points to where such evidence is located.

For multiple shapelet lengths $\{L_r\}$, we concatenate responses from all lengths to form the node-level pattern feature:

$$\mathbf{h}_i = \mathrm{Concat}(\mathbf{h}_i^{(L_1)}, \ldots, \mathbf{h}_i^{(L_R)}) \in \mathbb{R}^{D_s}, \tag{9}$$

where $D_s = \sum_r P_{L_r}$ is the total number of shapelet responses.

Localized responses can also be triggered by station-specific jitters, appearing as isolated activations without neighborhood support. We thus apply a lightweight one-hop refinement on the base graph to promote local spatial consistency:

$$\widetilde{\mathbf{H}} = \sigma\left(\widetilde{\mathbf{A}}_0\mathbf{H}\mathbf{W}_1\right), \tag{10}$$

where $\mathbf{H} \in \mathbb{R}^{N\times D_s}$ stacks $\mathbf{h}_i$ over all nodes; $\widetilde{\mathbf{A}}_0$ is the normalized base adjacency; $\mathbf{W}_1$ is a learnable projection; and $\sigma(\cdot)$ is a non-linear activation. The refined tokens $\widetilde{\mathbf{H}}$ are then fed to the subsequent spatial enhancement module.

3) Dynamic Adjacency and Multi-Hop Propagation: The one-hop refinement enforces neighborhood consistency of the pattern tokens, but it still relies on the static topology and only captures nearby neighbors. In practice, dependencies among stations can vary with the sensing context, such as temporary traffic controls and evolving traffic regimes, and anomalous impacts may propagate beyond local neighborhoods. To capture such context-dependent interactions, we learn a residual dynamic adjacency [11] on top of the base graph $\tilde{\mathbf{A}}_0$, and perform multi-hop propagation over the updated graph.

We parameterize an unconstrained residual matrix $\mathbf{R}$ and bound its entry-wise scale by $\tanh(\cdot)$. To restrict learnable edges, we apply an adjacency mask $\mathbf{M} \in \{0,1\}^{N\times N}$:

$$\Delta\mathbf{A} = \mathbf{M} \odot (\alpha \cdot \tanh(\mathbf{R})), \tag{11}$$

where $\alpha > 0$ controls the residual amplitude and $\odot$ denotes element-wise multiplication. We then update the adjacency and project it to a valid space consistent with our implementation:

$$\mathbf{A} = \mathrm{RowNorm}(\Pi(\tilde{\mathbf{A}}_0 + \Delta\mathbf{A})). \tag{12}$$

Here $\Pi(\cdot)$ is a projection operator that enforces structural constraints on the raw matrix:

$$\Pi(\mathbf{X}) = \mathrm{DiagKeep}_{\tilde{\mathbf{A}}_0}(\mathrm{Sym}(\mathrm{Clip}_{[0,1]}(\mathbf{X}))), \tag{13}$$

Algorithm 1: End-to-End One-Class Training and QoS-aware Inference

Input: Pure-normal training set $\mathcal{D}_{tr}$; normal validation set $\mathcal{D}_{val}$; test set $\mathcal{D}_{te}$; base adjacency A_0; tail length m; Top-k parameter k (or ratio ρ); target FPR α; loss weights $(\lambda_b, \lambda_t, \lambda_h)$; adjacency regularizers (λ_1, λ_2).

Output: Trained parameters $(\theta, \phi, \Delta A)$; threshold τ; alarms on $\mathcal{D}_{te}$.

```
Initialize (θ, φ) and ΔA ← 0;
for epoch ← 1 to E do
    Sample mini-batch {(X_t, e_t)} from D_tr;
    X̂_t, L_adj ← FORWARDAE(X_t, e_t; A_0, ΔA);
    Compute L_body, L_tail and L_hard via
        TAILRAMPMSE(X_t, X̂_t; m) and
        HARDTOPK(X_t, X̂_t; m, k);
    L ← λ_b L_body + λ_t L_tail + λ_h L_hard + L_adj;
    Update (θ, φ, ΔA) by backpropagation;
Compute validation node errors {r(X)} on D_val and
    fit a robust normalizer;
Compute validation scores {s(X)} on D_val by
    TAILTOPK(r(X); k);
Set threshold τ ← Q_{1−α}({s(X)});
foreach (X_t, e_t) ∈ D_te do
    X̂_t ← FORWARDAE(X_t, e_t; A_0, ΔA);
    s_t ← TAILTOPK(NODEERROR(X_t, X̂_t; m); k);
    if s_t > τ then
        Raise Alarm
```

where $\mathrm{Clip}_{[0,1]}(\cdot)$ clamps entries to $[0,1]$, $\mathrm{Sym}(\mathbf{X}) = (\mathbf{X} + \mathbf{X}^\top)/2$ symmetrizes the matrix to obtain an undirected adjacency, and $\mathrm{DiagKeep}_{\tilde{\mathbf{A}}_0}(\cdot)$ preserves the diagonal of $\tilde{\mathbf{A}}_0$ to retain self-connectivity. Finally, $\mathrm{RowNorm}(\cdot)$ performs row-wise normalization so that $\mathbf{A}$ is row-stochastic, which is convenient for random-walk style propagation.

Since $\mathbf{A}$ is data-driven, we regularize both the residual strength and the smoothness of node representations on the learned graph. Let $\mathbf{Z}^{(0)} = \tilde{\mathbf{H}}$ be the input token features to this spatial module, thus $\mathbf{Z}^{(b)}$ denotes the tokens of the b-th sample. We define:

$$\mathcal{L}_{\mathrm{adj}} = \lambda_1 \|\Delta \mathbf{A}\|_1 + \lambda_2 \cdot \frac{1}{B} \sum_{b=1}^{B} \mathrm{tr}\Big((\mathbf{Z}^{(b)})^\top \mathbf{L}(\mathbf{A}) \mathbf{Z}^{(b)} \Big), \quad (14)$$

where $\|\Delta \mathbf{A}\|_1$ discourages dense residual edges, $\mathbf{L}(\mathbf{A}) = \mathbf{D} - \mathbf{A}$ is the graph Laplacian, and the trace term encourages neighboring nodes to have smooth representations under the learned connectivity.

Finally, starting from $\mathbf{Z}^{(0)} = \tilde{\mathbf{H}}$, we perform K-hop propagation on $\mathbf{A}$ to capture longer-range spatial effects:

$$\mathbf{H}_S = \mathrm{MultiHopGCN}(\tilde{\mathbf{H}}; \mathbf{A}), \quad (15)$$

where $\mathbf{H}_S$ denotes spatially enhanced features. The resulting $\mathbf{H}_S$ is then passed to the fusion stage.

4) Cross-View Attention and Residual Field Fusion: The temporal tokens $\tilde{\mathbf{H}}$ capture localized shapelet-consistent precursor patterns, while the spatial tokens $\mathbf{H}_S$ encode context-dependent propagation on the learned graph. Since these two views emphasize different information, directly concatenating or averaging them may fail to exploit cross-view dependencies, especially when the spatiotemporal coupling varies across contexts. We therefore use cross-attention to align the temporal and spatial token sets and to exchange information between them, and further introduce a time-conditioned residual neural field [24] to adapt the fused representation to regime shifts.

To condition fusion on time, we construct a normalized time coordinate c for each sample and obtain its positional encoding:

$$c = \begin{cases} \dfrac{\mathrm{time\_step}}{T_{\mathrm{res}}}, & \text{if time\_step is provided,} \\ 1/T_{\mathrm{res}} \sum_{t=1}^{T_{\mathrm{res}}} \omega_t, & \text{otherwise,} \end{cases} \quad (16)$$
$$\mathbf{e}_t = \mathrm{PE}(c),$$

where T_{res} denotes the temporal resolution of the window. When an explicit time-step index is available, we normalize it to $c \in [0,1]$; otherwise, we use the normalized cumulative time coordinate with learnable increments $\{\omega_t\}$. $\mathrm{PE}(\cdot)$ is the positional encoder realized by index-based sinusoidal embeddings. We then apply cross-view attention to obtain the aligned representations:

$$\bar{\mathbf{H}}_T, \ \bar{\mathbf{H}}_S = \mathrm{CrossAttn}(\mathbf{H}_T, \mathbf{H}_S), \quad \mathbf{H}_T = \tilde{\mathbf{H}}, \quad (17)$$

where $\bar{\mathbf{H}}_T$ and $\bar{\mathbf{H}}_S$ are the attention-enhanced temporal and spatial tokens. We form a base fused content by averaging the two aligned views and generate a time-dependent residual correction via a residual neural field:

$$\mathbf{B} = (\bar{\mathbf{H}}_T + \bar{\mathbf{H}}_S)/2, \quad \Delta \mathbf{B}_t = \mathcal{F}_{\mathrm{RNF}}(\mathbf{B}, \mathbf{e}_t). \quad (18)$$

Here $\mathcal{F}_{\mathrm{RNF}}(\cdot)$ learns a time-conditioned residual mapping that adjusts the fused content according to the current temporal context. Since the correction strength should be content-adaptive, we predict a sample-wise gate:

$$\gamma = \sigma(\mathrm{MLP}(\mathrm{Pool}(\mathbf{B}))), \quad \mathbf{B}' = \mathbf{B} + \gamma \cdot \Delta \mathbf{B}_t, \quad (19)$$

where $\mathrm{Pool}(\cdot)$ is average pooling over the node dimension, $\gamma \in \mathbb{R}^{B \times 1}$ is broadcast to all nodes, and $\sigma(\cdot)$ is the sigmoid function.

Finally, we fuse the cross-attended tokens with the time-corrected fused features through a lightweight fusion network:

$$\mathbf{H}_F = \phi\big([\bar{\mathbf{H}}_T; \mathbf{B}']\big), \quad (20)$$

where $[\cdot;\cdot]$ concatenates along the feature dimension and $\phi(\cdot)$ is a two-layer MLP with dropout and normalization. Note that $\bar{\mathbf{H}}_T$ already incorporates spatial information through cross-attention, and $\mathbf{B}'$ is computed from both aligned views. The fused tokens $\mathbf{H}_F$ are fed to the decoder to reconstruct the spatiotemporal signals.

C. Decoder

Given the encoder output $\mathbf{z}_t \in \mathbb{R}^{B\times d_z}$ and the learned dynamic adjacency $\mathbf{A} \in \mathbb{R}^{N\times N}$, the decoder reconstructs the input window $\mathbf{X}_t$ by progressively mapping the latent code to node-wise hidden tokens, restoring spatial dependencies via graph decoding, and generating multivariate temporal signals for each node. To inject the time context during reconstruction, we concatenate $\mathbf{e}_t$ with the latent code:

$$\tilde{\mathbf{z}}_t = [\mathbf{z}_t; \mathbf{e}_t]. \tag{21}$$

Here $[\cdot;\cdot]$ denotes concatenation along the feature dimension, $\mathbf{B}$ is the batch size, and $\mathbf{e}_t \in \mathbb{R}^{B\times d_t}$ is the time-context embedding.

We first project $\tilde{\mathbf{z}}_t$ to node-wise hidden tokens and reshape them as:

$$\mathbf{U}_t = \mathrm{MLP}_u(\tilde{\mathbf{z}}_t), \tag{22}$$

where $\mathrm{MLP}_u(\cdot)$ outputs a vector in $\mathbb{R}^{B\times(Nd_u)}$, and $\mathbf{U}_t \in \mathbb{R}^{B\times N\times d_u}$ stacks d_u-dimensional tokens for all N nodes.

To restore spatial correlations in the reconstructed signals, we apply a multi-hop graph decoding operator over $\mathbf{A}$:

$$\mathbf{S}_t = \mathrm{MultiHopGCN}(\mathbf{U}_t; \mathbf{A}), \tag{23}$$

where $\mathbf{S}_t \in \mathbb{R}^{B\times N\times d_s}$ denotes the spatially decoded representations.

Finally, a temporal decoding head maps $\mathbf{S}_t$ to the multivariate sequence of length T for each node and reshapes it back to the window:

$$\hat{\mathbf{X}}_t = \mathrm{MLP}_t(\mathbf{S}_t), \tag{24}$$

where $\mathrm{MLP}_t(\cdot)$ outputs $T\cdot C$ values per node, yielding a tensor in $\mathbb{R}^{B\times N\times(TC)}$, and $\mathrm{Reshape}(\cdot)$ rearranges it into $\hat{\mathbf{X}}_t \in \mathbb{R}^{B\times N\times T\times C}$.

D. Pure-Normal One-Class Training Objective

QoS-aware early warning requires a training objective that (i) is sensitive to deviations near the end of a window, where precursors typically emerge, and (ii) does not rely on anomaly labels while remaining robust to the fact that anomalous impacts may involve only a small subset of nodes. We therefore adopt a pure-normal one-class training protocol: the training set $\mathcal{D}_{\mathrm{tr}}$ contains only windows verified to be normal, and event annotations are used only to filter training samples rather than supervise the model, avoiding any label leakage.

For a training window $(\mathbf{X}_t, \mathbf{e}_t) \in \mathcal{D}_{\mathrm{tr}}$, the encoder–decoder produces $\widehat{\mathbf{X}}_t$. We first define the per-time reconstruction error aggregated over nodes and channels:

$$\varepsilon_{t,\tau} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\Big(\mathbf{X}_t(i,\tau,c) - \widehat{\mathbf{X}}_t(i,\tau,c)\Big)^2, \tag{25}$$

where $\tau \in \{1,\dots,T\}$ indexes time steps within the window.

To emphasize late-stage deviations without introducing supervision, we decompose the reconstruction objective into a body term and a tail term of length m. The body loss averages errors before the tail:

$$\mathcal{L}_{\mathrm{body}} = \frac{1}{T-m}\sum_{\tau=1}^{T-m}\varepsilon_{t,\tau}, \tag{26}$$

while the tail loss applies a normalized monotone ramp over the last m steps:

$$\mathcal{L}_{\mathrm{tail}} = \frac{1}{m}\sum\nolimits_{j=1}^{m} w_j\varepsilon_{t,T-m+j}, \quad w_j = \frac{j}{\frac{1}{m}\sum_{\ell=1}^{m}\ell}. \tag{27}$$

This design reshapes the learning signal toward the window end while keeping the loss scale stable.

To focus optimization on a small set of high-error nodes, we introduce a Top-K focused training loss computed from node-wise tail errors:

$$r_{t,i} = \max_{\tau\in\{T-m+1,\dots,T\}}\frac{1}{C}\sum_{c=1}^{C}\Big(\mathbf{X}_t(i,\tau,c) - \widehat{\mathbf{X}}_t(i,\tau,c)\Big)^2, \tag{28}$$

and form a hard Top-K aggregation loss:

$$\mathcal{L}_{\mathrm{hard}} = \frac{1}{K}\sum\nolimits_{i\in\mathrm{TopK}(\{r_{t,i}\}_{i=1}^{N})} r_{t,i}. \tag{29}$$

Importantly, $\mathcal{L}_{\mathrm{hard}}$ is used only during training to focus gradients on a small set of high-error nodes, encouraging the model to fit normal patterns without being dominated by widespread benign fluctuations.

When the encoder learns a residual dynamic adjacency $\mathbf{A}$, we further regularize the learned graph to avoid overly dense connections and to promote smooth representations under the learned topology:

$$\mathcal{L}_{\mathrm{adj}} = \lambda_1\|\Delta\mathbf{A}\|_1 + \lambda_2\,\Omega(\mathbf{Z},\mathbf{A}), \tag{30}$$

where $\Delta\mathbf{A}$ is constrained by the adjacency mask and $\Omega(\mathbf{Z},\mathbf{A})$ enforces Laplacian smoothness of node features $\mathbf{Z}$ on $\mathbf{A}$. The overall objective on the pure-normal set is:

$$\mathcal{L}_{\mathrm{train}} = \lambda_b\mathcal{L}_{\mathrm{body}} + \lambda_t\mathcal{L}_{\mathrm{tail}} + \lambda_h\mathcal{L}_{\mathrm{hard}} + \mathcal{L}_{\mathrm{adj}}, \tag{31}$$

where $\lambda_b, \lambda_t, \lambda_h$ are non-negative hyperparameters controlling the contributions of each loss term, and we optimize $(\theta, \phi, \Delta\mathbf{A})$ by minimizing the expected loss over $\mathcal{D}_{\mathrm{tr}}$:

$$\min_{\theta,\phi,\Delta\mathbf{A}}\ \mathbb{E}_{(\mathbf{X}_t,\mathbf{e}_t)\sim\mathcal{D}_{\mathrm{tr}}}\big[\mathcal{L}_{\mathrm{train}}\big]. \tag{32}$$

E. Tail-Emphasized Top-K Scoring

After training on the pure-normal set $\mathcal{D}_{\mathrm{tr}}$, the model is deployed as an online early-warning detector that must operate under a target FPR budget α. In contrast to the training objective that shapes model parameters on normal data, inference follows a fixed scoring-and-thresholding rule: given a test window $(\mathbf{X}_t, \mathbf{e}_t)$, the model produces $\widehat{\mathbf{X}}_t$, computes a scalar score S_t, and raises an alarm if $S_t > \tau_\alpha$, where τ_α is calibrated on normal validation data to satisfy the QoS budget.

We start from the element-wise squared reconstruction error:

$$\mathbf{E}_t(i,\tau,c) = \Big(\mathbf{X}_t(i,\tau,c) - \widehat{\mathbf{X}}_t(i,\tau,c)\Big)^2. \tag{33}$$

TABLE I: Overall discrimination (AUC) and reconstruction quality (MSE) on three datasets.

Dataset	Metric	STG-RGCN	STG-GAT	GCN-LSTM	GCN	Transformer	MLP	CCB-GraphGAN	Our
FT-AED	MSE	0.0102	0.0089	0.0119	0.0095	0.0312	0.0122	**0.0061**	0.0082
	AUC	0.67	0.68	0.65	0.70	0.60	0.62	0.73	**0.75**
PeMS	MSE	0.0093	0.0098	0.0107	0.0112	0.0203	0.0126	0.0821	**0.0021**
	AUC	0.54	0.56	0.52	0.52	0.55	0.58	0.54	**0.85**
XTraffic	MSE	0.0487	0.0495	0.0541	0.0528	0.0632	0.0516	**0.0467**	0.0944
	AUC	0.59	0.58	0.57	0.54	0.56	0.53	0.62	**0.79**

TABLE II: QoS-aware early detection on FT-AED under multiple FPR budgets.

Model	Miss Percentage					Reporting Delay				
	20%	10%	5%	2.5%	1%	20%	10%	5%	2.5%	1%
STG-RGCN	8.33%	8.33%	16.67%	33.33%	**33.33%**	-11.55 ± 5.28	-7.18 ± 9.62	-8.95 ± 7.59	-5.81 ± 8.62	-1.56 ± 8.83
STG-GAT	**0%**	**0%**	16.67%	25%	41.67%	-14.04 ± 2.14	**-13.33 ± 2.56**	-7.75 ± 6.58	-6.11 ± 8.81	-1.85 ± 10.01
GCN-LSTM	**0%**	8.33%	25%	25%	41.67%	-10.08 ± 7.2	-8.18 ± 7.73	-6.83 ± 8.19	-5.11 ± 9.94	-2.29 ± 8.11
GCN	**0%**	16.67%	25%	25%	41.67%	-10 ± 8.05	-9.85 ± 6.50	-9.78 ± 6.89	-6.67 ± 9.17	-3 ± 9.56
Transformer	**0%**	25%	41.67%	66.67%	75.00%	-7.25 ± 6.36	-5 ± 6.25	-2.43 ± 8.35	-1.13 ± 8.34	+0.16 ± 7.31
MLP	25%	25%	41.67%	58.33%	58.33%	-12.33 ± 4.99	-7.44 ± 6.38	-6.86 ± 6.87	-8.5 ± 5.71	-2.8 ± 8.68
CCB-GraphGAN	**0%**	**0%**	8.33%	25%	**33.33%**	-14.08 ± 1.24	-10.58 ± 7.31	-7.77 ± 9.18	-8.22 ± 8.83	-5.06 ± 9.61
Our	**0%**	**0%**	**0%**	**8.33%**	**33.33%**	**-14.28 ± 3.35**	-11.02 ± 5.98	**-9.92 ± 6.90**	**-8.75 ± 4.51**	**-7.75 ± 4.51**

To emphasize early-warning sensitivity, we restrict scoring to the tail interval of length m, where precursors are more likely to emerge near the window end. We first aggregate errors over channels to obtain the per-node, per-time tail error:

$$\eta_{t,i,\tau} = \frac{1}{C}\sum_{c=1}^{C} \mathbf{E}_t(i,\tau,c), \quad \tau \in \{T-m+1,\dots,T\}. \tag{34}$$

We then summarize each node by its strongest deviation within the tail:

$$s_{t,i} = \max_{\tau\in\{T-m+1,\dots,T\}} \eta_{t,i,\tau}. \tag{35}$$

Since an anomalous event may affect only a small subset of locations, we form a window-level score by aggregating only the Top-K node scores:

$$S_t = \frac{1}{K}\sum_{i\in \mathrm{TopK}(\{s_{t,i}\}_{i=1}^{N})} s_{t,i}, \tag{36}$$

where $\mathrm{TopK}(\cdot)$ returns the indices of the largest K node scores, and S_t is the final score for window $\mathbf{X}_t$.

To meet the target FPR α, we calibrate the decision threshold using the normal validation set $\mathcal{D}_{\mathrm{val}}$ that follows the same pure-normal protocol:

$$\tau_\alpha = \mathrm{Quantile}_{1-\alpha}(\{S_t : (\mathbf{X}_t, \mathbf{e}_t) \in \mathcal{D}_{\mathrm{val}}\}), \tag{37}$$

where $\mathrm{Quantile}_{1-\alpha}(\cdot)$ denotes the $(1-\alpha)$-quantile computed on normal scores. An alarm is raised if $S_t > \tau_\alpha$. This calibration sets the operational low-FPR point directly from normal validation scores and makes the deployed detector compliant with the desired QoS budget.

V. Performance Evaluation

A. Setting

1) Datasets: We conduct experiments on three datasets.

FT-AED [12] is collected on the Interstate 24 near Nashville, Tennessee. We use the lane-level subset with 196 nodes, where speed, occupancy, and volume data are recorded every 30 seconds.

PeMS [25] is collected from the California highway system. We use the I-405 Southbound subset, which includes 114 monitoring stations and the lane-level measurements aggregated at a 5-minute interval.

XTraffic [26] is built from the California highway system. We use the subset with 196 stations, and adopt lane-level measurements aggregated at a 5-minute interval.

2) Data Preprocessing: For all datasets, we construct lane-level spatiotemporal graphs, where each node is a lane at a station with three variables: speed, occupancy, and volume. We segment the multivariate streams into overlapping sliding windows with length W and stride S.

To enforce pure-normal one-class training, we follow the event-filtering assumptions in [12] to remove potentially contaminated periods from the training set. Specifically, for FT-AED and PeMS, we allow up to 30-min reporting delay and exclude an additional 2-hour impact window. For FT-AED manual anomalies, we assume no delay but still exclude 2 hours after onset. For XTraffic, we use a more conservative 90-min delay and 4-hour impact window.

We apply chronological splits to avoid temporal leakage. For FT-AED, we train on 14 days, reserve 5 days for testing, and exclude one day due to excessive anomalies; the test period contains incidents of interest. For PeMS, we use the first 70%

TABLE III: QoS-aware cross-dataset performance of our method under FPR budgets.

Dataset	Miss Percentage					Reporting Delay				
	20%	10%	5%	2.5%	1%	20%	10%	5%	2.5%	1%
PeMS	1.20%	2.40%	4.79%	8.98%	11.38%	-21.58 $\pm$ 10.74	-17.36 $\pm$ 13.52	-15.79 $\pm$ 13.98	-14.93 $\pm$ 13.94	-13.92 $\pm$ 14.89
XTraffic	7.96%	17.13%	27.68%	41.70%	48.10%	-20.86 $\pm$ 12.57	-17.11 $\pm$ 13.86	-11.56 $\pm$ 15.84	-8.46 $\pm$ 6.75	-4.78 $\pm$ 3.86

TABLE IV: Ablation study under the 5% FPR budget.

Method	Metric	FT-AED	PeMS	XTraffic
w/o Dynamic Adj	AUC	0.61	0.79	0.76
	Miss	25%	25.15%	29.37%
	Delay	-8.83 $\pm$ 4.56	-12.48 $\pm$ 15.45	-9.65 $\pm$ 16.45
w/o MultiHop	AUC	0.63	0.83	0.78
	Miss	33.33%	10.18%	32.07%
	Delay	-9.17 $\pm$ 4.84	-15.17 $\pm$ 13.74	-10.17 $\pm$ 15.93
w/o Tail	AUC	0.71	0.81	0.76
	Miss	25%	24.55%	30.72%
	Delay	-8.28 $\pm$ 6.49	-13.96 $\pm$ 13.81	-10.96 $\pm$ 16.90
w/o Top-K	AUC	0.71	0.78	0.75
	Miss	33.33%	12.57%	29.85%
	Delay	-9.25 $\pm$ 4.39	-14.64 $\pm$ 13.4	-9.22 $\pm$ 15.279
Our	AUC	**0.75**	**0.85**	**0.79**
	Miss	**0%**	**4.79%**	**27.68%**
	Delay	**-9.92 $\pm$ 6.90**	**-15.79 $\pm$ 13.98**	**-11.56 $\pm$ 15.84**

for training and the remaining 30% for testing. For XTraffic, we train on the first 10 months and test on the last 2 months.

3) Metrics: To quantify and compare the anomaly detection performance of the models, we use the following metrics:

Miss Percentage measures the percentage of crash events that the model failed to detect within a detection window relative to their official reporting time (lower is better).

Reporting Delay quantifies detection time relative to the official report (lower is better). It is defined as the mean and standard deviation of the time difference between the detected time t_d and the report time t_r:

$$\text{Reporting Delay} = \mu(t_d - t_r) \pm \sigma(t_d - t_r). \tag{38}$$

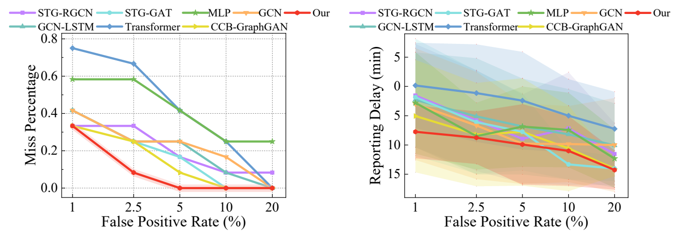
A negative delay indicates earlier-than-report detection.

Reconstruction Error measures the mean squared error (MSE) on normal traffic windows (lower is better).

AUC is the area under the Receiver Operating Characteristic (ROC) curve, measuring overall separability across all thresholds (higher is better).

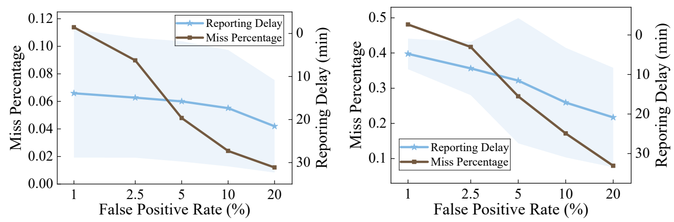
4) Baselines: We compare our method with six representative baselines, including spatiotemporal graph autoencoders: STG-RGCN AE [27] and STG-GAT AE [28], a spatial graph autoencoder: GCN AE [29], a spatial–temporal hybrid autoencoder: GCN-LSTM AE [30], a temporal-only autoencoder: Transformer AE [31], and a GAN-based lane-level detector: CCB-GraphGAN [15] that learns normal patterns via cycle-consistent bidirectional adversarial training and performs detection based on reconstruction errors.

5) Experimental Settings: All experiments are conducted under PyTorch and accelerated using a single NVIDIA GeForce RTX 3090 GPU. The batch size is set to 16, and the



(a) Miss Percentage (b) Reporting Delay

Fig. 3: QoS-aware early detection under FPR budgets.



(a) PeMS (b) XTraffic

Fig. 4: Cross-dataset early detection.

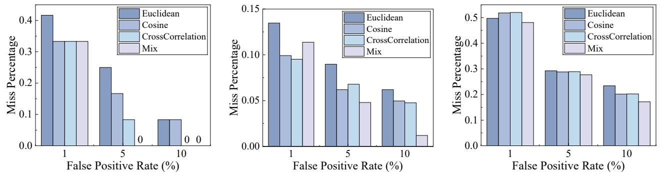
number of training epochs is set to 80 across all datasets. For FT-AED, we set W = 60, S = 5, and the shapelet configuration is {(10, 16), (15, 12), (20, 8), (25, 6)} (length, count). For PeMS, we set W = 24, S = 2, and the shapelet configuration is {(6, 16), (9, 12), (12, 8), (15, 6)}. For XTraffic, we set W = 48, S = 1, and the shapelet configuration is {(4, 16), (6, 12), (8, 8), (10, 6), (15, 4)}.

B. Overall Performance Comparison

1) Overall AUC and reconstruction quality: TABLE I summarizes AUC and MSE on three datasets. Our method achieves the best AUC on all datasets, indicating stronger separability between normal and incident-related traffic. In terms of MSE, our model is competitive on FT-AED and PeMS, reflecting effective modeling of normal patterns. Notably, on XTraffic, several baselines obtain lower MSE but much lower AUC, implying that low reconstruction error does not guarantee reliable separability under strict FPR-constrained early detection.

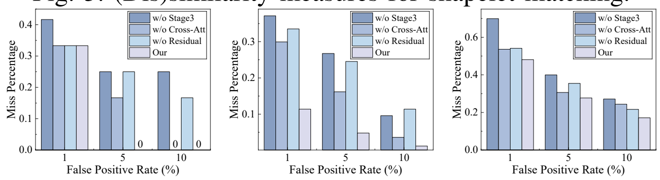
2) QoS-aware early detection under FPR budgets: We evaluate early detection on FT-AED across FPR budgets using Miss Percentage and Reporting Delay within a $\pm$15-minute window around a crash event. TABLE II shows 0% miss rate for our method at FPR $\geq$ 5%. At 1% FPR, our miss rate is comparable to the strongest baselines while detecting earlier with lower variance. Fig. 3 further shows that higher FPR budgets reduce misses for all methods, while our approach maintains a consistently favorable delay profile.

3) Cross-dataset early detection: On PeMS and XTraffic, TABLE I shows that most baselines yield near-random AUC,



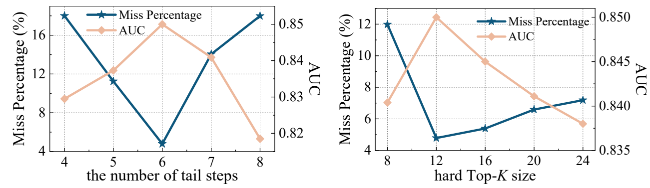
(a) FT-AED (b) PeMS (c) XTraffic

Fig. 5: (Dis)similarity measures for shapelet matching.

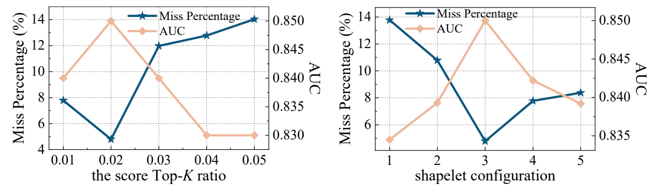


(a) FT-AED (b) PeMS (c) XTraffic

Fig. 6: Ablation study on cross-view attention and fusion.



(a) T_{tail} (b) K_{hard}



(c) ρ (d) $\mathcal{S}$

Fig. 7: Hyperparameter study on the PeMS dataset.

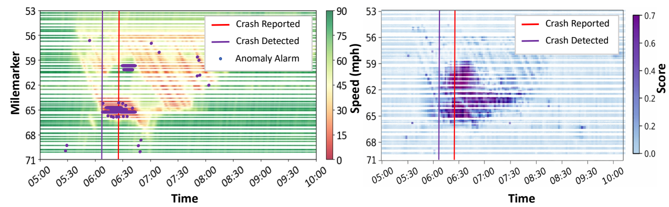


Fig. 8: Early-warning footprint of the crash event.

suggesting limited separability for incident-aware early detection in these settings. Therefore, as shown in TABLE III, we focus on the QoS-aware results of our method. For PeMS, we use a $\pm$15-minute window, and Fig. 4a shows consistently low miss rates with negative reporting delays. For XTraffic, we adopt a $\pm$30-minute window due to 5-min sampling and timestamp uncertainty in multi-source alignment. As shown in Fig. 4b, our method exhibits stable gains as the FPR budget increases, while triggering ahead of official reports. Overall, the results demonstrate robust early-warning performance across heterogeneous freeway networks under practical FPR budgets.

C. Ablation Study

We validate the necessity of our key designs and report AUC and the QoS-aware early-detection at 5% FPR.

1) Dynamic Adjacency and Multi-Hop Propagation: TABLE IV shows that replacing the dynamic adjacency with a static graph degrades performance across datasets and removing multi-hop propagation sharply worsens early detection, especially on FT-AED and XTraffic, highlighting the role of adaptive structure and diffusion for anomaly propagation.

2) Tail-Emphasized and Hard Top-K training: TABLE IV also validates that removing tail emphasis or hard Top-K mining noticeably increases missed events, indicating these QoS-oriented components improve early warning by focusing learning on tail deviations and hard abnormal segments.

3) Shapelet Matching Measures: Fig. 5 compares the proposed Mix metric with single-measure variants (Euclidean, cosine, cross-correlation) and shows consistently lower miss rates for Mix, supporting the benefit of combining complementary measures for robust shapelet representations.

4) Cross-View Attention and Residual Field Fusion: Fig. 6 demonstrates the necessity of the Stage-3 fusion design: removing Stage-3 causes the most severe degradation, and using only cross-view attention or only residual field fusion remains less effective than the full model. This confirms that both alignment and residual fusion are necessary for early warning.

D. Hyperparameter Analysis

We analyze four key hyperparameters on PeMS under FPR=5%: (1) T_{tail}, tail steps used in tail-emphasized scoring to highlight late-stage deviations; (2) K_{hard}, hard Top-K size for hard-sample mining during training; (3) ρ, Top-K ratio for forming QoS-aware scores; (4) the shapelet configuration $\mathcal{S}$, the lengths and counts of shapelets. The results are shown in Fig. 7. When T_{tail} is too small, delayed incident impacts may be under-emphasized; when it is too large, tail emphasis is weakened and performance degrades. A moderate K_{hard} provides sufficient hard-sample focus, while overly small or large values reduce stability. For ρ, selecting too few segments can miss informative deviations, whereas too many introduce noise and hurt separability. The shapelet configuration also shows a trade-off: too little capacity underfits local patterns, while too much adds cost with diminishing returns. We adopt the best-performing settings reported in the experimental setup.

E. Case Study

1) Crash early warning on FT-AED: For a representative crash in the FT-AED validation set, Fig. 8 (left) shows a localized speed breakdown and its spatiotemporal propagation. Under 10% FPR, our method triggers before the official report and keeps detections concentrated around the affected region. Fig. 8 (right) visualizes the decision score map, where high-score areas align with the degradation pattern, offering an intuitive corridor-level explanation.

2) Decision with Learned Shapelets: Fig. 9 shows the learned shapelets and their best-matching subsequences on normal and crash-related station series. Normal stations exhibit close matches and low scores, while crash-related stations show pronounced mismatches, indicating deviations from learned normal prototypes.

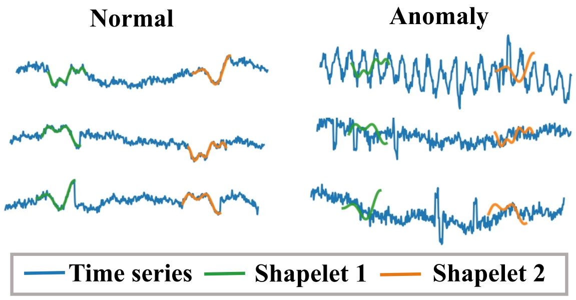


Fig. 9: Shapelet-matching evidence for the detection decision.

VI. Conclusion

In this paper, we introduce SE-STGAE, an early anomaly detection framework designed to address the challenges posed by weak and propagating precursors under strict false-alarm budgets in freeway sensing networks. We extract segment-level precursors with a shapelet-enhanced encoder and integrate them with spatiotemporal representations. Then, we learn propagation-aware spatial dependencies with adaptive adjacency and multi-hop diffusion, and fuse temporalspatial patterns to improve robustness to normal fluctuations. In practice, our study focuses on reconstruction-based early warning for freeway monitoring with offline calibration. Our future work will explore online adaptation under evolving traffic regimes and more robust calibration strategies.

Acknowledgment

This work is supported in part by National Natural Science Foundation of China under Grant Nos. 92567204, 62272193, and 62472194, and the Fundamental Research Funds for the Central Universities, and Jilin Science and Technology Research Project 20260203130SF, and Graduate Innovation Fund of Jilin University 2026CX206. Wenbin Liu is the corresponding author of this paper.

References

[1] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep learning for time series anomaly detection: A survey," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–42, 2024.

[2] S. Zhang, C. Zhou, Y. Liu, P. Zhang, X. Lin, and S. Pan, "Conformal anomaly detection in event sequences," in *Forty-second International Conference on Machine Learning*, 2025.

[3] B. M. Williams and A. Guin, "Traffic management center use of incident detection algorithms: Findings of a nationwide survey," *IEEE Transactions on intelligent transportation systems*, vol. 8, no. 2, pp. 351–358, 2007.

[4] S. Kim and B. Coifman, "Assessing the performance of speedinfo radar traffic sensors," *Journal of Intelligent Transportation Systems*, vol. 21, no. 3, pp. 179–189, 2017.

[5] P. Mercader and J. Haddad, "Automatic incident detection on freeways based on bluetooth traffic monitoring," *Accident Analysis & Prevention*, vol. 146, p. 105703, 2020.

[6] P. Zhang, W. Yi, Y. Song, P. Yan, P. Wu, A. Shemery, K. Hampson, and A. P. Chan, "Dynamic spatiotemporal graph network for traffic accident risk prediction," *GIScience & Remote Sensing*, vol. 62, no. 1, p. 2514330, 2025.

[7] Y. Sun, Y.-C. Lu, K. Fu, F. Chen, and C.-T. Lu, "Detecting anomalous traffic behaviors with seasonal deep kalman filter graph convolutional neural networks," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 4729–4742, 2022.

[8] C. Shyalika, H. K. Bagga, A. Bhatt, R. Prasad, A. A. Ghazo, and A. Sheth, "Time series foundational models: Their role in anomaly detection and prediction," *arXiv preprint arXiv:2412.19286*, 2024.

[9] W. Liu, H. Du, E. Wang, D. Luan, B. Yang, Y. Yang, and J. Wu, "Non-aligned multi-scale data completion for sparse mobile crowdsensing," in *IEEE/ACM International Symposium on Quality of Service*, 2025.

[10] H. Gao, R. Jiang, Z. Dong, J. Deng, Y. Ma, and X. Song, "Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting," *arXiv preprint arXiv:2312.00516*, 2023.

[11] F. Li, J. Feng, H. Yan, G. Jin, F. Yang, F. Sun, D. Jin, and Y. Li, "Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 1, pp. 1–21, 2023.

[12] A. Coursey, J. Ji, M. Quiñones-Grueiro, W. Barbour, Y. Zhang, T. Derr, G. Biswas, and D. B. Work, "Ft-aed: Benchmark dataset for early freeway traffic anomalous event detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 15 526–15 549, 2024.

[13] Y. Hu, A. Qu, and D. Work, "Detecting extreme traffic events via a context augmented graph autoencoder," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 6, pp. 1–23, 2022.

[14] L. Deng, D. Lian, Z. Huang, and E. Chen, "Graph convolutional adversarial networks for spatiotemporal anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2416–2428, 2022.

[15] M. Nouri and M. Ilbeigi, "Early detection of freeway traffic anomalies with a cycle-consistent bidirectional generative adversarial network," *Authorea Preprints*, 2024.

[16] R. Kaur, K. Sridhar, S. Park, S. Jha, A. Roy, O. Sokolsky, and I. Lee, "Codit: Conformal out-of-distribution detection in time-series data," *arXiv preprint arXiv:2207.11769*, 2022.

[17] O. A. Ekle and W. Eberle, "Anomaly detection in dynamic graphs: A comprehensive survey," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 8, pp. 1–44, 2024.

[18] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 947–956.

[19] Z. Liang, J. Zhang, C. Liang, H. Wang, Z. Liang, and L. Pan, "A shapelet-based framework for unsupervised multivariate time series representation learning," *arXiv preprint arXiv:2305.18888*, 2023.

[20] N. Zhang and S. Sun, "Multiview unsupervised shapelet learning for multivariate time series clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4981–4996, 2022.

[21] X. Qiu, X. Wu, Y. Lin, C. Guo, J. Hu, and B. Yang, "Duet: Dual clustering enhanced multivariate time series forecasting," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2025, pp. 1185–1196.

[22] X. Wan, L. Cen, X. Chen, and Y. Xie, "Multi-generator adversarial dynamic spatial–temporal shapelet network for anode effect prediction in aluminum electrolysis process," *Advanced Engineering Informatics*, vol. 62, p. 102609, 2024.

[23] R. Younis, A. Hakmeh, and Z. Ahmadi, "Mts2graph: Interpretable multivariate time series classification with temporal evolving graphs," *Pattern Recognition*, vol. 152, p. 110486, 2024.

[24] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.

[25] S. Guo, Y. Lin, H. Wan, X. Li, and G. Cong, "Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 11, pp. 5415–5428, 2021.

[26] X. Gou, Z. Li, T. Lan, J. Lin, Z. Li, B. Zhao, C. Zhang, D. Wang, and X. Zhang, "Xtraffic: A dataset where traffic meets incidents with explainability and more," *arXiv preprint arXiv:2407.11477*, 2024.

[27] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[29] T. Kipf, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[30] A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.