

GroundSTory: Interactive Storytelling for Grounded Spatiotemporal Anomaly Explanations with Graphs and Large Language Models

JUANPABLO HEREDIA and LEIGHTON ESTRADA-RAYME, Fundação Getulio Vargas (FGV), Brazil

JEREMY MATOS-CANGALAYA, University of Engineering and Technology (UTEC), Peru

JORGE POCO, Fundação Getulio Vargas (FGV), Brazil

Explaining anomalies in spatiotemporal data requires more than identifying statistical deviations; it often demands coherent narratives that relate events to their spatial, temporal, and social context. In urban settings, where anomalies may reflect public events, behavioral shifts, or environmental factors, the lack of interpretable, contextual explanations limits the effectiveness of automated analysis. We present GROUNDSTORY, an interactive framework that reframes spatiotemporal anomaly explanation as computational storytelling, integrating graph-based representations of urban space, coordinated visual analytics, and large language models (LLMs) to support the construction of grounded, human-readable narrative explanations. Anomalies act as narrative anchors, while localized spatial and temporal relations and contextual signals derived from the graph guide and constrain narrative generation. The LLM functions as a narrative assistant, proposing plausible, evidence-backed hypotheses that analysts can inspect, compare, and validate through visual interaction. We illustrate the approach through qualitative usage scenarios on real-world urban datasets from São Paulo and New York City, showing how interactive narratives surface meaningful links between detected anomalies and contextual factors such as holidays, mass events, and routine activity patterns. A prompt analysis further suggests that structured, graph-derived context tends to affect the stability and perceived quality of generated hypotheses. Overall, our results indicate that combining knowledge-driven representations, interactive visualization, and constrained prompting can support more coherent, plausible, and interpretable narratives for spatiotemporal anomaly analysis, contributing to ongoing research in interactive storytelling and narrative-centered AI systems.

CCS Concepts: • **Human-centered computing** → **Visualization systems and tools**; **Visual analytics**; • **Computing methodologies** → *Natural language generation*; • **Information systems** → *Spatial-temporal systems*.

Additional Key Words and Phrases: Spatiotemporal anomaly analysis, interactive explanation, graph-based modeling, large language models, visual analytics, urban data analytics

ACM Reference Format:

Juanpablo Heredia, Leighton Estrada-Rayme, Jeremy Matos-Cangalaya, and Jorge Poco. 2026. GroundSTory: Interactive Storytelling for Grounded Spatiotemporal Anomaly Explanations with Graphs and Large Language Models. *ACM Trans. Intell. Syst. Technol.* 1, 1 (January 2026), 25 pages. <https://doi.org/10.1145/3856820>

Authors' Contact Information: Juanpablo Heredia, juanpablo.parillo@fgv.br; Leighton Estrada-Rayme, leighton.rayme@fgv.br, School of Applied Mathematics, Fundação Getulio Vargas (FGV), Rio de Janeiro, RJ, Brazil; Jeremy Matos-Cangalaya, jeremy.matos@utec.edu.pe, University of Engineering and Technology (UTEC), Lima, Peru; Jorge Poco, jorge.poco@fgv.br, School of Applied Mathematics, Fundação Getulio Vargas (FGV), Rio de Janeiro, RJ, Brazil.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2026 Copyright held by the owner/author(s).

ACM 2157-6912/2026/1-ART

<https://doi.org/10.1145/3856820>

1 Introduction

Spatiotemporal (ST) data are increasingly available across domains such as urban monitoring, mobility analysis, environmental sensing, and public safety. These data streams capture complex dynamics unfolding across space and time and are routinely analyzed for tasks such as forecasting, classification, imputation, and anomaly detection [21]. Among these, anomaly detection plays a critical role by highlighting unusual or potentially consequential events. However, most existing anomaly detection methods operate as black boxes: they flag deviations without articulating the surrounding context, underlying factors, or plausible interpretations that might explain why an anomaly occurred [6, 31]. This limitation is particularly acute in high-stakes urban settings, where anomalies often intersect with social dynamics, public events, or shifts in human behavior. In such contexts, understanding an anomaly requires more than numerical justification — it requires a coherent narrative that situates the event within its spatial, temporal, and social environment. While prior systems offer visual summaries or statistical indicators, few support the construction of human-readable, contextualized explanations that combine spatial structure, temporal evolution, and verifiable contextual evidence in a unified, interactive manner [13, 33].

Recent advances in large language models (LLMs) have renewed interest in narrative generation and computational storytelling, owing to their ability to synthesize information into fluent natural-language accounts [41]. At the same time, LLMs are known to suffer from hallucination, lack of grounding, and limited explainability when used in isolation [4, 41]. These challenges have motivated growing research on integrating LLMs with symbolic representations, such as graphs and structured data, to constrain generation, improve interpretability, and support interactive sensemaking [4, 16, 41]. Within this emerging landscape, explanation can be reinterpreted not as a static output, but as a narrative artifact that evolves through interaction between computational models and human analysts.

This article extends our *SIBGRAPI 2025* paper [17], which introduced the underlying visualization-based framework for ST anomaly analysis. The present version expands this foundation through a broader conceptual framing, additional empirical evaluations, and a deeper analysis of grounded narrative generation. Building on these extensions, we introduce *GROUNDSTORY*, a framework that conceptualizes ST anomaly explanation as an interactive narrative construction process by combining graph-based modeling, interactive visual analytics, and LLM-driven narrative generation grounded in localized spatial and temporal context (Fig. 1). The framework couples anomaly selection with graph-derived context retrieval and constrained prompting to produce grounded, inspectable narratives. Rather than operating as an autonomous decision-maker, the LLM serves as a narrative assistant, proposing data-supported hypotheses that analysts can inspect, compare, and iteratively refine through visual interaction.

The graph representation plays a dual role: it encodes spatial structure and temporal variation, and it provides a computational scaffold for grounding narratives. Nodes represent urban entities associated with time series, while edges capture geographic proximity. Coordinated visualizations enable users to explore anomalies across multiple scales and select focal events for analysis. The graph thus acts as a constraint that grounds and bounds narrative generation: structured prompts derived from graph structure and local temporal context steer the LLM toward explanations tied to verifiable spatial and temporal evidence. This coupling—reframing anomaly explanation as graph-grounded, interactive storytelling—is the conceptual core of our contribution, positioning explanation as a human-in-the-loop sensemaking process rather than a one-shot textual output.

We demonstrate the utility of this approach in urban monitoring, where interpretability and contextual understanding are essential. Using real-world datasets from São Paulo and New York City, we show how interactive narrative explanations can surface plausible connections between

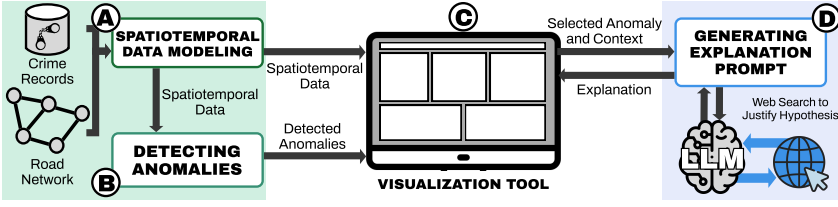


Fig. 1. GROUNDSTORY pipeline. (A) Model ST data from incident records and the urban road network. (B) Detect anomalies in the ST data. (C) Support interactive exploration via coordinated visualizations. (D) Retrieve the local context of a selected anomaly and generate grounded LLM explanations. © 2025 IEEE. Reprinted, with permission, from Heredia *et al.*, *Proc. SIBGRAPI 2025* [17].

detected anomalies and contextual factors such as holidays, mass events, or changes in routine activity. Our evaluation combines qualitative usage scenarios with an exploratory prompt analysis examining how structured context influences perceived narrative quality and grounding.

In summary, our contributions are threefold:

- GROUNDSTORY, a narrative-driven framework that reframes ST anomaly explanation as interactive computational storytelling, using the graph both to ground narratives in coordinated visualizations and to constrain LLM generation toward verifiable spatial and temporal evidence.
- An interactive visual analytics system that integrates spatial, temporal, and narrative views to support exploratory sensemaking and hypothesis refinement.
- A multi-city empirical evaluation demonstrating how grounded LLM-generated narratives aid anomaly interpretation in urban data.

Together, these contributions establish anomaly explanation as a grounded, interactive storytelling process, positioning interpretability and analyst control at the center of urban anomaly analysis.

2 Related Work

Our work lies at the intersection of three research threads: ST anomaly detection methods for urban data, which provide the computational backbone for identifying candidate events; grounding large language models with graphs and structured context to improve controllability and reduce hallucinations; and urban visual analytics systems that support interactive exploration and sensemaking. We review these areas and highlight how our framework combines them to enable narrative-centered, human-in-the-loop anomaly explanation.

2.1 Spatiotemporal Anomaly Detection in Urban Contexts

Identifying anomalies requires first defining normal behavior. In urban settings, behaviors considered abnormal in one location or period may be typical in another [39]. ST patterns in cities — such as traffic flows or crowd dynamics — are shaped by both routine and irregular events. This variability demands context-aware approaches to anomaly detection [39, 40]. Anomaly detection is commonly tackled using machine learning or statistical methods, often through reconstruction- or forecasting-based strategies that identify deviations from expected patterns. In what follows, we group prior work into three method families commonly used in urban ST settings and discuss their implications for interpretability and downstream explanation.

Statistical and classical methods. These methods model normal ST behavior using explicit assumptions about distributions, linearity, or low-rank structure. Spatial scan statistics and local

indicators (e.g., Local Moran's I) detect localized spatial deviations but typically rely on simplified temporal windows. Decomposition-based approaches (e.g., PCA, NMF, tensor factorization) represent normal patterns via low-rank components and identify anomalies as sparse residuals [34, 36]. Autoregressive models such as STAR and STARMA capture local ST dependencies for forecasting-based anomaly detection [2, 25, 28]. Overall, these methods are typically unsupervised, computationally efficient, and interpretable, and can be effective with limited data; however, they are sensitive to modeling assumptions and may fail to capture complex, nonlinear dynamics [2, 6].

Graph-based methods. These methods model urban data as spatial graph signals, leveraging spectral representations to detect localized or multi-scale irregularities. Graph wavelets and Laplacian-based filtering highlight anomalies as deviations in transformed graph-frequency domains [7, 37]. Hybrid approaches integrate graph priors into classical low-rank and sparse models, preserving geometry while isolating anomalous behavior. For instance, GLOSS introduces Laplacian regularization into tensor decomposition, bridging algebraic methods with graph-structured assumptions [34].

Deep learning methods. These methods learn ST representations directly from data. Standard RNNs and CNNs capture temporal dynamics, whereas graph neural networks explicitly model spatial adjacency and neighborhood interactions. Many approaches frame anomaly detection as prediction or reconstruction error, including graph-based autoencoders and adversarial models such as MTGAE [29], STADN [35], and STGAN [8]. Hybrid GNN-RNN architectures dominate recent work, offering high expressive power and state-of-the-art accuracy [29, 35], but they typically require large training datasets and careful tuning, and offer reduced interpretability.

2.2 LLMs and Urban Data Graphs

Recent studies have adapted LLMs to address urban data problems, where ST dependencies and limited labeled data pose unique challenges [20, 24]. These models aim to generalize across diverse tasks, including traffic forecasting, crime prediction, and policy planning.

Graph-based methods can ground LLMs in relational urban context, improving relevance and reducing redundancy [26]. A common approach linearizes graph elements (nodes, edges, or paths) into text that an LLM can ingest, though performance is sensitive to encoding choices and task design [10]. Complementing this, Perozzi *et al.* [27] introduce a GNN encoder that maps graph information into the LLM token space, enabling tighter alignment. Such techniques often pair with knowledge-graph retrieval via indexed subgraph reconstruction [16, 19]. In contrast, our system eschews global indexing and retrieves only the local ST context around a user-selected node.

In the urban domain, UrbanGPT [24] and UrbanLLM [20] illustrate the promise of ST-aware LLMs. UrbanGPT embeds spatial and temporal dependencies for zero-shot prediction, while UrbanLLM decomposes tasks and routes subtasks to specialized models. Inspired by these, we embed spatial and temporal features into a structured prompt that guides the LLM through step-by-step reasoning to generate context-aware hypotheses.

More broadly, these grounded LLM-graph integrations align with narrative-centered AI: they support the generation of compact, checkable explanations that connect observed deviations to contextual cues, and they enable users to iteratively refine and validate emerging hypotheses. This perspective motivates our design choice to treat anomaly explanation as interactive computational storytelling rather than one-shot summarization.

2.3 Urban Data Visualization and Interactive Analysis

Interactive visual analytics is key to deciphering complex ST patterns in urban data. Cities continually generate heterogeneous, high-velocity streams, and carefully designed visualizations help uncover correlations, trends, and anomalies across space and time. Deng *et al.* [9] emphasize that

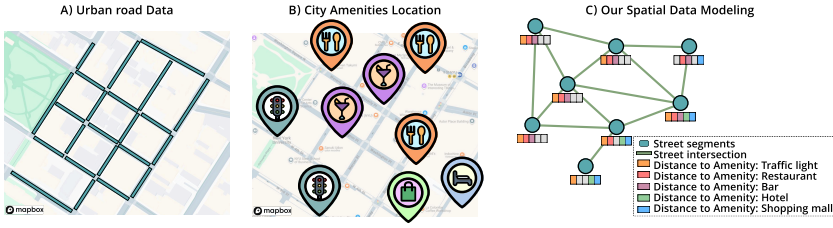


Fig. 2. Spatial data modeling. (A–B) Raw urban data. (C) Graph construction with OSMnx and node grounding using amenity distances.

exploring urban causal relationships requires tools capable of handling noise, temporal variation, and spatial dependence.

Several systems target pattern discovery and decision support in ST contexts through coordinated visualizations and specialized techniques such as ST-Heatmaps [30]. A common focus is on human-computer interaction and interpretability [12, 30]. For example, *CrimAnalyzer* [12] pairs NMF-based hotspot detection with linked views to examine crime dynamics, while *JamVis* [30] aids congestion analysis of Waze alerts through clustering and visual summarization. Going beyond descriptive pattern identification, *Compass* [9] embeds causal inference by extending Granger causality to reveal dynamic relationships in urban time series, accompanied by tailored visual explanations.

Similarly, our work aims to empower users through interactive exploration and augments these visual methods with LLM-guided reasoning, thereby fusing visual and linguistic interpretability within a single framework.

3 Data and Anomaly Detection Backbone

This section describes the computational backbone that supports our interactive explanation workflow. We first define a graph-based representation for urban ST data, then introduce the families of anomaly detection methods considered in our framework, and finally instantiate one representative detector per family. Sec. 4 builds on the anomalies produced here to construct grounded narrative explanations through human-in-the-loop interaction.

3.1 Spatiotemporal Data Modeling

Following Hassan *et al.* [15], we begin by selecting a study area in the target city and extracting its road network using OSMnx [5], resulting in an intersection-centric graph where nodes represent street intersections and edges represent street segments. Because our analysis focuses on street segments, we invert this representation: every original edge becomes a node, and two nodes are linked when their corresponding street segments intersect (see Fig. 2A and C).

The graph is enriched with three types of node attributes: *time series*, *amenities*, and *street characteristics*. Each node corresponds to a street segment and stores a multivariate time series capturing incident counts. For each city, we aggregate incident records into temporal bins and represent each street segment as a time series of incident counts over the corresponding period. To incorporate contextual information, we also attach amenity profiles to each node by querying OpenStreetMap (OSM) points of interest within a buffer around the corresponding street segment. Amenities are grouped into categories (e.g., restaurants, schools, nightlife) and encoded as distance vectors to each category. Finally, street characteristics, like segment length and road type, are extracted from OSM metadata.

To support method selection and downstream explanation, we treat this graph as the shared computational substrate: anomaly detectors operate over its node-level time series, while the explanation component later retrieves localized neighborhoods and contextual signals from the same representation.

3.2 Detecting Anomalies

Our framework is compatible with a range of ST anomaly detection methods. Since cities and datasets may favor different inductive biases, we consider representative families of approaches—statistical/classical, graph-based, and deep learning methods—as interchangeable backbones within a common interactive explanation workflow.

To concretely instantiate these families, we select one representative detector from each category: a classical forecasting baseline (STARMA), a graph spectral approach (SGWT), and a deep learning model (STADN). This choice allows us to examine how different inductive biases surface different anomaly types (e.g., localized spikes versus sustained shifts) while keeping the downstream explanation pipeline unchanged.

3.2.1 Statistical and Classical Methods: STARMA. The Space–Time Autoregressive Moving Average (STARMA) model extends ARMA/ARIMA to capture both temporal and spatial dependencies in multivariate time series [28]. It forecasts normal behavior and flags points deviating beyond a threshold. For N spatial locations observed at discrete time points, the STARMA(p, q, r, s) formulation is

$$\mathbf{Y}_t = \underbrace{\sum_{i=1}^p \sum_{j=0}^r \boldsymbol{\phi}_{ij} \mathbf{W}^{(j)} \mathbf{Y}_{t-i}}_{\text{Autoregression (AR)}} + \underbrace{\sum_{i=1}^q \sum_{j=0}^s \boldsymbol{\theta}_{ij} \mathbf{W}^{(j)} \boldsymbol{\varepsilon}_{t-i}}_{\text{Moving Average (MA)}} + \boldsymbol{\varepsilon}_t. \quad (1)$$

Here, $\mathbf{Y}_t \in \mathbb{R}^N$ is the vector of observations at time t ; p and q denote the temporal AR and MA orders, respectively; r and s specify the spatial AR and MA orders—that is, how many neighborhood “rings” are considered at each lag; $\mathbf{W}^{(j)} \in \mathbb{R}^{N \times N}$ is the j -th-order spatial weight matrix (here, derived from the graph adjacency); $\boldsymbol{\phi}_{ij}$ and $\boldsymbol{\theta}_{ij}$ are the AR and MA coefficient matrices for temporal lag i and spatial lag j ; and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ represents white-noise innovations. In this process, the current observation $\mathbf{Y}_t$ depends on past values (AR term) and past disturbances (MA term), not only from the same location but also from neighboring areas.

Model fitting. Coefficients $\{\boldsymbol{\phi}_{ij}, \boldsymbol{\theta}_{ij}\}$ are estimated by maximum likelihood. Hyperparameters (p, q, r, s) are selected via a grid search guided by the Akaike and Bayesian Information Criteria (AIC/BIC).

Anomaly detection. After fitting, the model yields one-step-ahead forecasts $\hat{\mathbf{Y}}_t$. We flag an observation as anomalous when any component of the residual $\mathbf{r}_t = \mathbf{Y}_t - \hat{\mathbf{Y}}_t$ exceeds ± 3 standard deviations of the in-sample residual distribution. Since STARMA embeds both spatial contiguity and temporal dynamics, this residual-based rule highlights localized deviations that purely temporal or purely spatial models might miss.

3.2.2 Graph-Based Methods: SGWT. We adopt the Spectral Graph Wavelet Transform (SGWT) introduced by Hammond *et al.* [14], which defines wavelets through spectral filtering on the graph Laplacian. Let $G = (V, E)$ be an undirected weighted graph with combinatorial Laplacian $L = D - W$ and eigendecomposition $L = U\Lambda U^\top$, where $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{N-1})$ with $0 = \lambda_0 \leq \dots \leq \lambda_{N-1}$. For a graph signal $f \in \mathbb{R}^N$, the Graph Fourier Transform is $\hat{f}(\lambda_\ell) = u_\ell^\top f$, and spectral filtering is

defined as

$$g(L)f = U g(\Lambda) U^\top f = \sum_{\ell=0}^{N-1} g(\lambda_\ell) \widehat{f}(\lambda_\ell) u_\ell.$$

Wavelets are defined by a band-pass kernel $g(\lambda)$ satisfying $g(0) = 0$, ensuring localization of high-frequency content. Scaling is performed spectrally: for scale $\tau > 0$,

$$W_\tau f = g(\tau L)f = \sum_{\ell=0}^{N-1} g(\tau \lambda_\ell) \widehat{f}(\lambda_\ell) u_\ell.$$

Small scales τ emphasize high graph frequencies (large eigenvalues), producing wavelets highly localized around nodes; large scales emphasize low frequencies, capturing more global variations.

The spectral kernel $g(\lambda)$ controls spatial localization via the decay properties of $g(\tau \lambda)$. Because $g(\tau \lambda_\ell)$ selectively amplifies specific frequency bands, each wavelet coefficient $(W_\tau f)_i$ represents the signal variation around node i at resolution τ . High-frequency components detect sharp local irregularities (e.g., discontinuities, anomalies, boundary effects), while low-frequency responses reflect large-scale smooth structure aligned with graph connectivity.

Model fitting. A graph signal assigns a scalar value to each node. To model ST data with N nodes and T time steps, we construct a temporal graph by stacking the spatial graph T times and adding inter-layer edges that connect each node to itself at adjacent time steps. We then apply the SGWT using the standard pygsp implementation with an Abspline filter bank, scaled by a factor of four. Graph filtering is approximated using Chebyshev polynomials following Hammond *et al.* [14], enabling efficient spectral filtering without explicit eigendecomposition.

Anomaly detection. From the SGWT coefficients, we extract, for each node, the response in the highest-frequency band, which captures localized and abrupt signal variations. After taking absolute values and standardizing the coefficients, anomalies are identified as nodes whose responses exceed ± 3 standard deviations, adopting the conventional three-sigma criterion, applied uniformly across detector families for comparability.

3.2.3 Deep Learning Methods: STADN. The Spatial-Temporal Anomaly Detection Network (STADN) [35] is a deep learning framework for anomaly detection in multivariate time series generated by interconnected sensors or nodes. It follows a predictability-based paradigm, where anomalies are identified as deviations from learned normal ST dynamics. STADN explicitly models both temporal dependencies within each time series and spatial dependencies across sensors, enabling joint detection and localization of anomalous behavior.

STADN consists of two main components. The **Predictive Module** (PM) captures spatial dependencies using a Graph Attention Network (GAT) and temporal dependencies using a Long Short-Term Memory (LSTM) network. Sensor relationships are represented as a directed graph, where attention weights determine the relative influence of neighboring sensors. For each node, the GAT aggregates information from its neighbors, while the LSTM encodes short- and long-term temporal patterns from a sliding window of historical observations. The resulting spatial and temporal representations are concatenated and passed through fully connected layers to predict future values for each sensor. The **Error Refinement Module** (ERM) then reconstructs the prediction errors produced by the PM and adds a learned refinement term to the initial predictions, amplifying the separation between normal and anomalous patterns.

Model fitting. STADN is trained in an unsupervised or semi-supervised manner using data assumed to represent normal system behavior. During training, input data are preprocessed to mitigate the influence of extreme outliers (by normalizing points with large variance), allowing the model to focus on steady-state dynamics. The Predictive Module is first optimized by minimizing

the mean squared error (MSE) between predicted and observed sensor values over a sliding temporal window. After convergence, the PM parameters are fixed, and the Error Refinement Module is trained separately to reconstruct and enhance prediction errors. This two-stage training strategy prevents the model from relying excessively on the refinement module while preserving the PM's predictive capacity.

Anomaly detection. At inference time, anomalies are identified based on prediction errors, defined as the absolute difference between predicted and observed values at each node and time step. Prediction errors are robustly normalized across sensors using median and interquartile range statistics to account for scale differences. A global anomaly score for each timestamp is computed by taking the maximum normalized error across all sensors. A timestamp is flagged as anomalous when this score exceeds a predefined threshold, while the contributing sensor-level errors enable spatial localization of anomalies.

3.3 Comparing Methods

We compare these methods qualitatively and quantitatively, not to advance detection performance, but to select an interpretable backbone for the explanation workflow.

Qualitative comparison. Table 1 summarizes strengths and limitations that matter in our setting. Beyond these general trade-offs, we emphasize three qualitative aspects that directly affect downstream explanation and interactive storytelling. First, *localization and scale*: STARMA tends to highlight temporally sharp deviations at specific nodes when residuals spike, whereas SGWT can expose multi-scale spatial irregularities through high-frequency localized responses; STADN can capture more complex nonlinear patterns but may distribute errors across many nodes depending on training dynamics. Second, *interpretability of the anomaly signal*: STARMA provides parametric interpretability through explicit space-time lag coefficients, while SGWT offers signal-based

Table 1. Comparison between anomaly detection methods.

Method	Strengths	Limitations
STARMA	High Parametric Interpretability: Coefficients directly quantify space-time lagged influence of neighboring nodes on a variable's value.	High Model Complexity: Resource-intensive parameter identification. Assumptions: Relies on strong assumptions, which may not hold in practice.
SGWT	Efficiency and Multi-scale Decomposition: Acts as a linear filter, enabling multiresolution analysis (low-frequency global trends vs. high-frequency local variations). Training-Free: No explicit training.	Threshold Sensitivity: Performance highly dependent on user-defined energy threshold. Expertise Required: Needs domain expertise to design the underlying graph kernel function.
STADN	Superior Latent Modeling: Uses Graph Attention Networks and LSTMs to capture complex, non-linear ST dependencies. High Customizability: Adaptable via custom loss functions for specific objectives.	Model Opacity: Low interpretability, inherent to deep learning (black-box). High Computational Burden: Resource-intensive training. Strict Preprocessing: Sensitive to data quality; requires strict normalization to stabilize training.

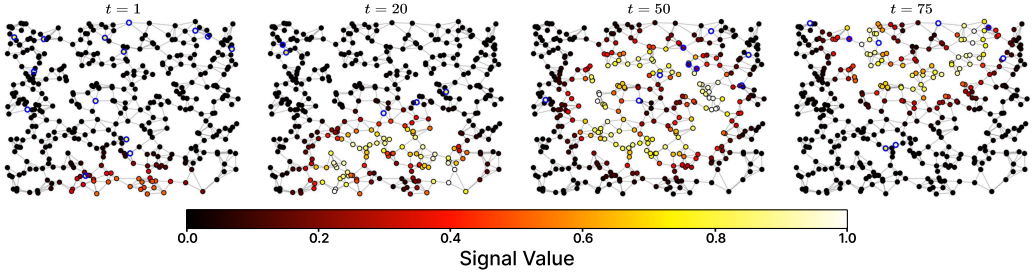


Fig. 3. Synthetic spatiotemporal data. A static random graph receives a Gaussian ring signal moving from bottom to top; anomalies are injected as random peaks or valleys (highlighted in blue).

interpretability by attributing anomalies to localized variations in the neighborhood-scale graph-frequency response. In contrast, STADN typically yields anomaly scores that are harder to trace back to specific mechanisms, which can complicate human validation and narrative refinement. Third, *controllability and stability*: STARMA and SGWT expose relatively transparent knobs (model orders and residual thresholds; kernel/scale choices and coefficient thresholds), enabling analysts to reason about why a point was flagged and to compare detectors consistently across cities. These qualitative considerations motivate our focus on STARMA and SGWT as primary backbones for the interactive explanation stage.

Quantitative comparison (synthetic data). To measure performance, we construct a synthetic dataset (a random graph with a synthetic signal) and inject anomalies at random, as shown in Fig. 3 (754 injected anomalies over 400 nodes and 100 timesteps; see the supplementary Appendix A for synthetic data generation). Because ground-truth anomaly labels are available only in this synthetic setting, we report standard classification metrics here, summarized in Table 2. STARMA performs best on most metrics, while the SGWT-based method is competitive.

Detector agreement (real data). For the São Paulo crime dataset, no ground-truth anomaly labels exist, so we do *not* report performance metrics for it; instead, we examine the agreement among detectors purely to characterize the difficulty and ambiguity of anomaly detection on real urban data. STARMA and SGWT share more anomalies (around 40%) than all three methods together (around 12%); STADN produces the highest proportion of exclusive detections (45% of its flagged points), indicating substantially different behavior. Such agreement or disagreement does not imply correctness: concordant detectors may be jointly wrong, and divergent ones jointly informative.

Given its lower performance on the synthetic benchmark and reduced interpretability for the explanation workflow, we do not use STADN in the subsequent usage scenarios. STARMA and SGWT, by contrast, yield locally interpretable anomaly signals, capturing lagged space-time influence and neighborhood-scale graph-frequency variation, respectively.

Table 2. Results on synthetic data with classification metrics. (↑) Higher is better; (↓) lower is better.

Method	Precision ↑	Recall ↑	F1 score ↑	F2 score ↑	TP ↑	FP ↓	FN ↓
STARMA	<u>0.92</u>	0.81	0.86	0.83	612	<u>54</u>	142
SGWT	0.97	0.77	0.86	<u>0.81</u>	583	17	171
STADN	0.69	<u>0.78</u>	0.73	0.76	<u>589</u>	262	<u>165</u>

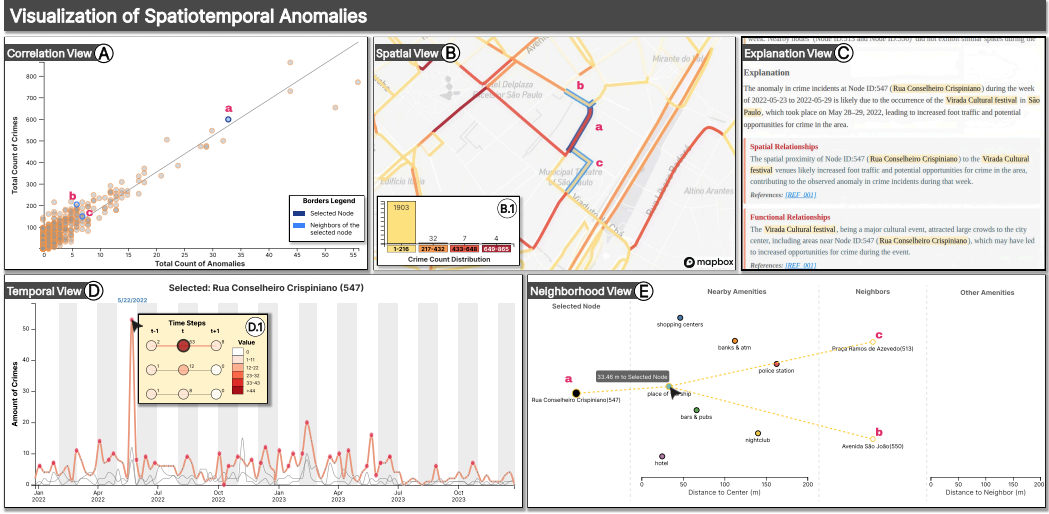


Fig. 4. The visualization tool. (A) Correlation View: incident and anomaly counts per node. (B) Spatial View: node map and selected neighborhood; colors encode incident counts (B.1). (C) Explanation View: LLM-generated narrative. (D) Temporal View: selected-node time series with anomaly context on hover (D.1). (E) Neighborhood View: selected-node local subgraph and nearby amenities. © 2025 IEEE. Reprinted, with permission, from Heredia *et al.*, *Proc. SIBGRAPI 2025* [17].

4 Interactive Storytelling for Grounded Explanation

We frame ST anomaly explanation as an interactive process of narrative construction. In our setting, anomalies serve as *narrative anchors*: user-selected focal events around which analysts assemble spatial, temporal, and contextual evidence into coherent accounts. This section presents the GROUNDSTORY interface and grounded LLM pipeline that support this human-in-the-loop storytelling workflow.

4.1 Visualization Tool

The visualization tool operationalizes our storytelling framing by coupling *selection*, *contextualization*, and *narration*. Analysts begin by identifying candidate anomalies, then inspect their spatial and temporal neighborhood to gather supporting cues, and finally request and refine grounded explanations in the narrative panel. The coordinated views share selection and filtering state, enabling users to iteratively move between anomaly anchors, local evidence, and generated hypotheses. In practice, analysts follow an iterative loop: select an anomalous event, inspect its local spatial and temporal context, generate one or more narrative hypotheses, validate them against contextual cues and references, and refine or compare explanations as needed.

Correlation View. This scatterplot maps each street-segment node to a point whose coordinates encode the total recorded incidents (e.g., crimes, traffic accidents) and the corresponding anomaly count (Fig. 4A). The view enables rapid detection of outliers and clusters, serving as a gateway to deeper exploration. Users can click points or brush across multiple ones to drive the *Spatial View*. To minimize visual clutter, we display only nodes with at least one recorded incident and provide pan-and-zoom controls for navigation. Point outlines convey selection state — orange by default, blue for the active node, and light blue for its immediate neighbors.

Spatial View. This map-based pane depicts the study area’s spatial domain (Fig. 4B). Each street-segment node appears as a selectable object, and the view reflects any selections made in the *Correlation* or *Temporal Views*—whether via brushing or time-range filtering—so analysts can focus on either the entire network or a chosen subset. Node fill color represents the total incident count, with progressively darker hues indicating higher values. Counts are divided into four equal-width intervals, shown in the legend (Fig. 4B.1). Opacity encodes anomaly status: nodes without detected anomalies or outside the active filters are desaturated. Clicking a node highlights it in **blue**, while its first-order neighbors appear in **light blue**. All selections are propagated back to the *Correlation View*, ensuring consistent encoding across the interface. Fig. 4 illustrates this interaction: the selected node is labeled **a**, and its neighbors are labeled **b** and **c**.

Explanation View. This panel presents LLM-generated explanations, formatted for clarity and linked to the visual interface (Fig. 4C). Hovering over a node ID in the text highlights the corresponding elements across views, connecting narrative and visualization. A drop-down menu allows analysts to switch between *GPT-4o Mini* and *Gemini 2.0 Flash*, both of which use a common prompt template. When web validation is enabled (Sec. 4.2), the system uses web-augmented configurations of these models to retrieve and cite external references; otherwise, the model is constrained to the provided ST context to generate context-only hypotheses.

Temporal View. When a node is chosen, its incident-count time series is plotted (Fig. 4D); otherwise, the network-wide aggregate series is shown. Moving the cursor reveals the exact date and public holidays. Neighboring nodes’ series appear as faint background traces to provide local context. Hovering over an anomaly marker (red dots) displays a tooltip with incident counts for the selected node and its neighbors in the previous, current, and subsequent intervals (Fig. 4D.1). Two interactions are available: (i) clicking an anomaly marker requests an LLM explanation by retrieving the associated subgraph and a temporal window of -4 to $+2$ intervals; (ii) holding Alt while clicking activates a time-range selection that syncs the *Correlation View* and *Spatial View*.

Neighborhood View. This panel organizes the spatial context of the selected node into four regions (Fig. 4E): (1) the selected node; (2) amenities within a fixed radius of that node; (3) first-order neighboring nodes; and (4) amenities that are close to at least one neighbor but not to the selected node. This view complements the *Spatial* and *Temporal Views* by presenting connectivity and proximity as spatial evidence for (or against) LLM-generated hypotheses. Hovering over an amenity reveals its distance and linkage (see cursor in Fig. 4E). Overall, the *Neighborhood View* provides an interactive summary of local contextual cues used to ground narrative explanations.

4.2 Prompting for Anomaly Explanation

Explanations are generated through a structured prompt (Fig. 5) that constrains the LLM to operate within the provided ST context while producing machine-parseable outputs that can be automatically checked for schema completeness (Fig. 6) and reference presence.

4.2.1 Prompt Structure. A short system message specifies the model’s role and global behavioral constraints (e.g., language, tone, and output format), including explicit prohibitions against speculative claims and format deviations.

The main prompt is composed of six components:

- (1) **Problem statement:** defines the task and provides a concise operational definition of an anomaly (e.g., “an unexpected deviation relative to spatial neighbors, temporal patterns, or both”).
- (2) **Task structure:** a multi-step procedure split into (i) hypothesis generation based solely on the ST context, (ii) web validation, and (iii) an XML report. Instructions are expressed in an imperative and rule-based form (e.g., “Generate at least two hypotheses”, “Do not use external

Problem Statement	An anomaly has been detected in a specific week and street; a spatiotemporal analysis is needed to identify possible causes and provide a plausible explanation. An anomaly is an unexpected value compared to spatial neighbors, temporal patterns, or both.
Task Structure	Complete these three sequential parts. Do not let later parts influence earlier analysis. Part 1: Initial Hypotheses (Data-Only Analysis) - Identify the anomaly's spatial extent and temporal duration based on the provided graph data. - Generate at least 2 distinct hypotheses explaining the indicated anomaly. - Base hypotheses exclusively on the graph data: amenities, crime patterns, holidays, and spatial relationships. - Consider why this specific location/time was affected differently than its direct spatial neighbors. Part 2: Web Validation - Crucial Geographic Focus: Search for verifiable evidence exclusively from {city_name}, focusing on events, news, or public safety reports that occurred specifically in the vicinity described by the graph during the anomalous period. - Prioritize information related to the specific streets mentioned in the graph, or the immediate surrounding blocks. Avoid general city-wide news or events that are not spatially specific to this small area. - Focus on: local news outlets, official municipal event calendars, public safety reports, and government announcements relevant to {city_name}. - Evaluate each hypothesis generated in Part 1 against the web evidence found. - Select the single best-supported hypothesis. - If no relevant web evidence is found specifically for the designated small geographic area and time, explicitly state this. - If the evidence found is not highly consistent with the precise spatial or temporal location (e.g., it's a city-wide event or in a different neighborhood), disregard it and continue searching for more specific evidence, or state that no relevant evidence was found. Part 3: Final XML Report Format your conclusion using the specified XML structure, including only web-verified sources in the References section.
Spatio-Temporal Context	Textual Graph: Contains nodes with street info, nearby amenities, crime time series, holiday markers, and an explicit anomaly indicator <pre> === Nodes === {Node ID}({node name}, {node type}) Nearby amenities (distances in meters): - {amenity}: {distance} ... Crime Incidents (Weekly): - {start week date} to {end week date}: {crimes count} {*(Anomaly)*} ... === Edges === {node name} ({node id}) -- {node name} ({node id}) ... </pre>
Extra Spatial Context	Amenity Security Risk Reference: - {amenity} : {short behavior description}. - Bars/Pubs: Alcohol-related violence, late hours - ...
Extra Temporal Context	Holiday list per week - {start week date} to {end week date} include {holidays on this week}
Output Structure	Final considerations - Always reference nodes by ID: (id: xxxx) - Use only real, accessible URLs in References - Base conclusions on verifiable evidence only - You must adhere strictly to the following output structure: <pre> <AnomalyExplanation> ... </AnomalyExplanation> </pre>

Fig. 5. Prompt structure for the São Paulo crime case. The left column lists prompt components and the right shows an abbreviated example. © 2025 IEEE. Adapted, with permission, from Heredia *et al.*, *Proc. SIBGRAPI 2025* [17].

knowledge in Part 1”), encouraging structured reasoning without requiring the model to output intermediate rationales.

- (3) **Textualized ST context:** a compact, graph-derived description of the anomalous street, including node identifiers, nearby amenities with distances, and localized time-series excerpts. Graph edges are represented only through node names and IDs to maintain compactness and avoid unnecessary structural complexity.
- (4) **Extra spatial context:** nearby amenities and short behavioral descriptors (e.g., alcohol-related activity or late-night operation), supporting location-based hypotheses.
- (5) **Extra temporal context:** holidays, notable events, or seasonal markers explicitly associated with the anomalous time window.
- (6) **Output structure (XML):** a strict, machine-readable structure (Fig. 6) that enforces separation between hypotheses, validated evidence, and references, and prevents free-form generation.

The prompt’s core implements a two-phase logic. First, the model proposes multiple *plausible* explanations based exclusively on the provided ST context, explicitly disallowing external knowledge or web sources. Second, in the web-validation phase, the model must retrieve verifiable evidence

```

<AnomalyExplanation>
  <Details>
    <Spatial>Describe affected nodes/areas.</Spatial>
    <Temporal>Start and end dates/times of anomaly.</Temporal>
    <Patterns>Key patterns observed in data.</Patterns>
  </Details>
  <Hypothesis>
    <Statement>State the chosen hypothesis clearly.</Statement>
    <Reasoning>
      <Relational>
        <Description>How spatial relationships support hypothesis.</Description>
        <Evidence>Comma separated list of references, e.g., REF_001,REF_002</Evidence>
      </Relational>
      <Functional>
        <Description>How functional relationships support hypothesis.</Description>
        <Evidence>Comma separated list of references.</Evidence>
      </Functional>
      <Temporal>
        <Description>How timing and duration support hypothesis.</Description>
        <Evidence>Comma separated list of references.</Evidence>
      </Temporal>
    </Reasoning>
  </Hypothesis>
  <Conclusion>Summary of findings and overall explanation.</Conclusion>
  <Keywords>Comma separated list of relevant terms.</Keywords>
  <References>
    <Item>
      <Id>Reference id, e.g., REF_001</Id>
      <Source>e.g., Local News Article Title</Source>
      <URL>https://example.com/news-article</URL>
      <Excerpt>Quote or summarize relevant part supporting claim.</Excerpt>
      <Date>Date of news/event</Date>
    </Item>
  </References>
</AnomalyExplanation>

```

Fig. 6. XML output schema for structuring hypotheses, evidence, and references. © 2025 IEEE. Reprinted, with permission, from Heredia *et al.*, *Proc. SIBGRAPI 2025* [17].

and assess whether each reference supports or contradicts the initial hypotheses. This separation preserves exploratory reasoning while enforcing grounded, source-cited conclusions.

4.2.2 LLM Constraints. The prompt design follows established practices to improve controllability and reduce ambiguous model behavior. First, tasks are presented as a sequence of explicit imperative instructions (e.g., “Select the single best-supported hypothesis”, “Disregard evidence that is not spatially specific”), a strategy associated with better instruction adherence and reduced speculative reasoning [23]. Second, the task decomposition mirrors principles of structured reasoning and chain-of-thought prompting [38], while avoiding unconstrained free-form rationales by restricting intermediate outputs. Finally, the output is constrained to a strict XML schema, following best-practice guidelines for safe, machine-readable generation [3].

Together, these constraints act as a risk-mitigation strategy: they encourage evidence-backed, geographically specific explanations while reducing hallucinations and maintaining consistency.

5 Implementation and Reproducibility

This section summarizes implementation details for the reproducibility of GROUNDSTORY, including system architecture, data-processing parameters, prompt artifacts, and runtime considerations.

5.1 System Architecture

The processing pipeline was developed in Python 3.9 using open-source packages such as NumPy, Pandas, GeoPandas, and Scikit-learn. Anomaly detection was executed in a Google Colab environment without GPU support. The visualization system follows a client-server paradigm. The server handles data filtering and selection, communicates with OpenAI and Gemini APIs, and parses and

processes their responses. The local server was implemented using Python's Flask package. For the client (visualization system), we used JavaScript with the Svelte framework. All visualizations were created using the *D3.js* library.

LLM requests and parsing. For each explanation request, the server assembles a prompt by serializing (i) the selected anomaly anchor (node and time point), (ii) localized spatial context (neighbor node IDs and amenity cues with distances), and (iii) localized temporal context (a short window around the selected time point, plus any available markers such as holidays). Model outputs are returned in a strict XML format (Sec. 4.2). The server parses the response into structured fields (hypotheses, evidence, references, etc.) and forwards them to the client for display and cross-view highlighting.

Model configurations. The system supports two modes: context-only hypothesis generation and an optional web-augmented validation mode (Sec. 4.2). In both, the client issues requests through the same interface, and the server selects the model configuration and attaches the corresponding prompt blocks. We tested two web-augmented configurations: *GPT-4o mini* (gpt-4o-mini-search-preview) with built-in search, and *Gemini 2* (gemini-2.0-flash) with the Google Search tool. *GPT-4o mini* tended to produce concise, well-structured hypotheses with cleanly formatted citations, whereas *Gemini 2* returned longer reference lists but matched evidence to claims less consistently and was less specific about locations and time periods.

5.2 Data Processing and Parameters

We follow the data modeling process described in Sec. 3.1, and we highlight here the parameters that most directly affect reproducibility and interpretation.

Temporal aggregation. Incident records are aggregated into temporal bins of size Δt (weekly and monthly bins in our experiments), producing one incident-count time series per street segment.

Neighborhood retrieval. Neighborhood context is computed on the road-network graph and retrieved on demand for interactive explanation. Unless otherwise stated, we use the 1-hop neighborhood of the selected street segment and a temporal window of -4 to $+2$ intervals around the selected time point.

Amenity extraction and categories. Amenities were queried from OpenStreetMap within a 200 m radius of each street segment and mapped into coarse categories (transport, food & nightlife, education, commerce, public services, and entertainment & lodging). The complete mapping of amenity types to categories is provided in the supplementary Appendix B.

Detector configuration. STARMA hyperparameters (p, q, r, s) are selected via AIC/BIC grid search (Sec. 3.2.1); SGWT uses a standard filter bank with Chebyshev approximation (Sec. 3.2.2); STADN follows the reference implementation unless explicitly stated (Sec. 3.2.3). Where applicable, we apply a consistent thresholding rule based on standardized residuals/coefficients (e.g., $\pm 3\sigma$) to facilitate comparison across detector families.

5.3 Prompt Templates and Schemas

To reduce uncontrolled model variation and support reproducibility, we use a fixed prompt template and a strict XML output schema (Fig. 6). The template follows the six-component structure described in Sec. 4.2 (problem statement, task structure, textualized ST context, extra spatial context, extra temporal context, and output structure). For further detail, examples of explanation prompts and responses are provided in the supplementary Appendices C.1 and C.3, respectively.

Schema checks. After generation, we parse the XML response and verify the presence of required fields (e.g., multiple hypotheses in phase one and evidence/references in the validation phase). If parsing fails or required fields are missing, the output is marked as invalid and, optionally, it is regenerated with a brief corrective instruction.

5.4 Runtime Considerations

Execution time. The anomaly detection procedure takes ~ 4.5 s per run (~ 4.0 s for STARMA fitting and ~ 0.5 s for residual computation and thresholding). LLM-based explanations require ~ 14 s per query, depending on service latency and context length. STARMA hyperparameter tuning takes ~ 1 h and is an offline, one-time step.

6 Prompt Ablation Study

To probe whether and how prompt configuration and graph-derived context affect hypothesis generation, we conduct a small-scale, controlled prompt-ablation study using an LLM-as-a-judge protocol. This exploratory study serves as a reference point to assess whether our prompt components (i) improve robustness to superficial perturbations and (ii) increase the perceived quality of hypotheses as a richer ST context is provided. We present this analysis as a validation of a core design choice in GROUNDSTORY: by comparing generation with and without the structured, graph-derived context, we assess whether these components are necessary to obtain robust, higher-quality explanations. We use three state-of-the-art models as evaluators: GPT-5 (gpt-5-mini), Gemini 2.5 (gemini-2.5-flash), and Claude Sonnet (claude-sonnet-4-20250514). By relying on an ensemble of judges, we reduce reliance on a single evaluator and mitigate the idiosyncratic preferences that can arise in LLM-based meta-evaluation. Examples of judge prompts and responses are provided in the supplementary Appendices C.2 and C.4, respectively. Our analysis is structured into two stages. First, we evaluate the effect of evidence-selection constraints on robustness and preference. Second, we examine how richer textualized graph context affects the sensitivity and perceived quality of generated hypotheses.

6.1 Evaluation Protocol and Metrics

We adopt a multi-layered evaluation framework inspired by recent meta-evaluation studies on LLM-as-a-judge reliability [11, 22]. Our evaluation focuses on two complementary properties of a judge model: (i) internal consistency under controlled perturbations and (ii) sensitivity to meaningful differences in contextual quality.

(1) Internal Consistency (Robustness). This dimension measures whether the judge produces stable decisions when the evaluated content remains semantically identical. We quantify this property using **Intra-Pair Instability (IPI)** [11], which captures the judge’s susceptibility to presentation noise.

For each evaluation case, candidate responses are presented multiple times with permuted ordering. A *flip* occurs when the judge selects a different response solely due to a change in presentation order. The IPI metric is defined as:

$$\text{IPI} = \frac{\text{Number of cases where the selected response changes across permutations}}{\text{Total number of unique evaluation cases}}$$

Lower values indicate greater robustness. Following Feng *et al.* [11], we interpret $\text{IPI} < 0.1$ as high consistency, indicating that fewer than 10% of decisions are order-dependent.

In scenarios involving more than two candidate responses, we additionally report a **Pairwise Consistency Rate**. This metric measures agreement between independent judging trials of the same response pair:

$$\text{Pairwise Consistency Rate} = \frac{\text{Number of matching pairwise decisions}}{\text{Total number of evaluated pairs}}$$

A consistency rate above 85% suggests low stochasticity and reliable local preferences.

(2) Discriminative Sensitivity. This dimension evaluates whether the judge can reliably distinguish responses of differing contextual richness. Inspired by the LaaJMeter framework [1], we assess sensitivity by progressively augmenting the amount of ST and contextual information provided to the model.

Let $\text{lvl } x$ denote a specific context level (e.g., minimal context, spatial context, or full ST context). Sensitivity at level x is defined as:

$$\text{Sensitivity}(\text{lvl } x) = \frac{\text{Number of trials in which responses at level } x \text{ are preferred}}{\text{Total number of trials}}$$

A monotonic increase in preference for higher context levels indicates that the judge meaningfully leverages additional information rather than relying on superficial cues.

6.2 Stage I: Sensitivity to Evidence Selection Constraints

The first stage measures sensitivity to the strict selection of web sources (specified in **Part 2** of the prompt’s *Task Structure*). This constraint enforces geographic focus (local sources) and temporal alignment (evidence matching the anomaly period). We operationalize this comparison as a binary decision task: judges choose between outputs generated *with* the evidence-selection constraints and outputs generated *without* them. This stage is based on a small set of *five* test cases, so we treat its results as indicative rather than conclusive.

Judges evaluate these responses along three axes: *relevance* (alignment with the anomaly), *credibility* (trustworthiness), and *diversity* (coverage of perspectives).

As shown in Table 3, Gemini exhibited perfect stability (IPI = 0.00, 95% CI [0.00, 0.43]), while Claude and GPT-5 showed moderate variability (IPI = 0.20 and 0.40, respectively). We report confidence intervals because IPI is estimated from a small number of evaluation cases (a single flip substantially changes the estimate). The 95% confidence interval provides a range of plausible values for the true instability rate under repeated sampling, making the uncertainty of the estimate explicit. The intervals are wide in our setting due to the small number of cases used. Positional bias was not statistically significant for any model, as the confidence intervals for first-option selection included 0.5, indicating that the observed differences are consistent with random variation.

6.3 Stage II: Impact of Textual Graph Richness

The second stage tests how the amount and structural richness of graph-derived context shape hypothesis quality. This stage uses a four-option comparison, a harder setting where judge consistency is expected to be lower—we therefore stress reliability and interpret the results as indicative trends. For each anomalous node, we construct four progressively enriched contextual views:

- **Level 1 (single_street):** core attributes of the selected street only.
- **Level 2 (single_street_ts):** street attributes augmented with its time-series excerpt and holiday markers.
- **Level 3 (street_neighbors):** Level 2 augmented with immediate neighbors and their attributes.

Table 3. Intra-Pair Instability (IPI) and positional bias for Stage I (binary choice).

Model	IPI	95% CI	Bias (First)
Gemini 2.5	0.00	[0.00, 0.43]	0.40 [0.23, 0.59]
Claude Sonnet	0.20	[0.04, 0.62]	0.44 [0.27, 0.63]
GPT-5	0.40	[0.12, 0.77]	0.44 [0.27, 0.63]

Table 4. Sensitivity to context (selection ratios and 95% confidence intervals) for Stage II.

Model	Lvl 1 (Low)	Lvl 2	Lvl 3	Lvl 4 (High)
Gemini 2.5	0.04 [0.01, 0.20]	0.08 [0.02, 0.25]	0.32 [0.17, 0.52]	0.56 [0.37, 0.73]
Claude Sonnet	0.08 [0.02, 0.25]	0.12 [0.04, 0.30]	0.28 [0.14, 0.48]	0.52 [0.33, 0.70]
GPT-5	0.04 [0.01, 0.20]	0.08 [0.02, 0.25]	0.40 [0.23, 0.59]	0.48 [0.30, 0.67]

- **Level 4 (all):** full configuration combining spatial, temporal, and attribute context for the street and its neighborhood.

Judges compare hypotheses according to three criteria: *spatiotemporal insight* (depth), *innovation* (novelty relative to alternatives), and *completeness* (coverage of relevant anomaly dimensions).

In this four-option setting, internal consistency decreased relative to the binary task. The Pairwise Consistency Rate dropped to a mean of approximately 0.64 (Gemini: 0.64, Claude: 0.60, GPT: 0.68). This decline suggests that as the number of competing alternatives increases, the judge's local preference stability becomes more stochastic. Despite this increase in variability, Table 4 shows a consistent trend toward deeper contextual information. The high-context condition (Level 4) was selected in 0.520 of trials, whereas Level 1 was selected in only 0.053 of trials. Aggregate preferences follow a clear gradient:

$$\text{Lvl 1 (0.053)} < \text{Lvl 2 (0.093)} < \text{Lvl 3 (0.333)} < \text{Lvl 4 (0.520)}.$$

Because consistency is moderate and the number of trials is limited, the confidence intervals are relatively wide and partially overlap, so the differences should be interpreted cautiously. Nevertheless, the overall ordering suggests that judges tend to prefer explanations generated with richer and more structured ST context.

7 Usage Scenarios

We evaluate our framework through two datasets and three usage scenarios. These datasets were selected because they (i) represent distinct types of urban risk, (ii) differ in temporal density and spatial heterogeneity, and (iii) allow us to demonstrate the generality of the method across cities, time granularities, and contextual constraints.

7.1 Datasets

Data sources. Crime occurrences were downloaded from the official public security portal of the State of São Paulo¹. Crash events were obtained from New York City's open data portal². For spatial amenities (Points of Interest, PoI), we rely on the OSMnx library, which provides a unified interface to OpenStreetMap categories.

São Paulo crime (weekly). We analyze official cell-phone theft records from a dense downtown region of São Paulo, following the spatial extent used by Hassan *et al.* [15]. This setting is challenging: some streets exhibit very high incident counts while others remain nearly inactive. We use data from 2022–2023 and aggregate occurrences weekly. For our experiments, we rely solely on approximate coordinates (latitude and longitude) and timestamps, omitting more granular metadata. Weekly aggregation balances temporal resolution and data sparsity, preserving local variation while ensuring sufficient support for ST modeling. For contextual grounding, the prompt incorporates (i) a list of official local holidays per week and (ii) PoI categories that are repeatedly associated with

¹<https://www.ssp.sp.gov.br/estatistica/consultas>

²<https://data.cityofnewyork.us/d/h9gi-nx95>

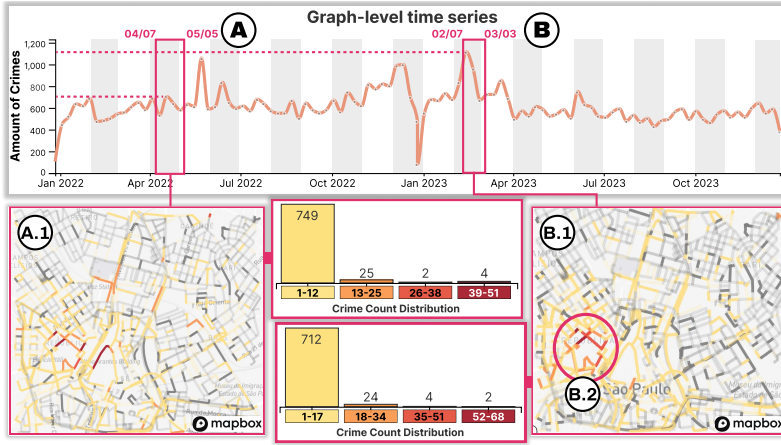


Fig. 7. Carnival-related periods from April 7 to May 5, 2022 (A), and February 7 to March 3, 2023 (B). (A.1) and (B.1) show the corresponding Spatial Views; (B.2) highlights the area with the highest incident concentration. © 2025 IEEE. Reprinted, with permission, from Heredia *et al.*, *Proc. SIBGRAPI 2025* [17].

opportunity structures and routine activities, including metro stations, shopping centers, banks and ATMs, police stations, places of worship, bars and pubs, and nightclubs. These cues provide structured signals that often co-vary with crime activity.

New York City crashes (monthly). To test generality across domains and geographies, we analyze traffic crash data from Manhattan during 2015–2024. This dataset differs substantially from São Paulo crime: crashes are widely distributed (nearly every street segment registers some incidents), but the per-location frequency is often low. Thus, we aggregate data monthly, which alleviates sparsity while retaining seasonal urban mobility patterns. Temporal grounding includes coarse weather descriptors, as meteorological conditions can influence crash risk. Based on historical NYC climate profiles, we annotate December with light snowfall; January and February with heavier snowfall; and March with end-of-season snow. These weather categories serve as coarse, manually curated descriptors designed to capture broad seasonal trends. We also annotate major events related to the COVID-19 pandemic (e.g., the initial outbreak period in New York City and the start of vaccination), which provide salient temporal context for interpreting large deviations. The PoI set is tailored to mobility and traffic characteristics, including parking areas, charging stations, bars, cinemas/theaters, and traffic signals.

7.2 Scenarios 1 and 2: São Paulo Crimes

We begin with the São Paulo cell-phone theft dataset to illustrate how our framework supports grounded explanations in a dense, heterogeneous urban core. Analysts often encounter both *positive* anomalies (sudden spikes) and *negative* anomalies (unexpected drops) whose interpretation depends on local routines, holidays, and event-driven disruptions. The two scenarios below demonstrate how interactive exploration and LLM-guided narrative generation help connect detected anomalies to plausible contextual drivers, while keeping hypotheses grounded in localized spatial cues and verifiable temporal references.

7.2.1 Carnival of 2023. Across Brazil, Carnival brings a surge in street activity — characterized by tourism, parades, increased alcohol consumption, and crowding — that can correlate with a rise in opportunistic crime. In 2022, lingering pandemic effects delayed celebrations to April 16–30,

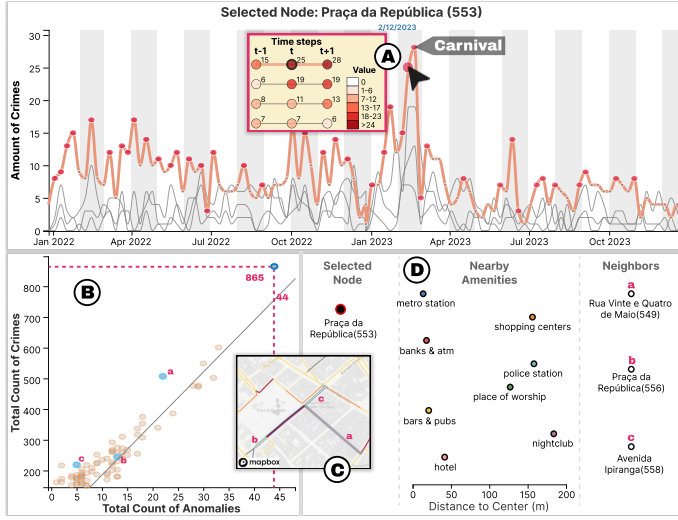


Fig. 8. High-incident street segment during the Carnival period. (A) Local context around the peak. (B) Correlation View with the selected node and neighbors. (C) Spatial View. (D) Neighborhood View with nearby context and amenities. © 2025 IEEE. Reprinted, with permission, from Heredia *et al.*, *Proc. SIBGRAPI 2025* [17].

while in 2023, they occurred on February 18–21. Fig. 7 (*Temporal View A and B*) marks the weeks surrounding both events.

We focus on node 553, near *Praça da República*, which recorded the highest counts in both years (see the spatial context in Fig. 7B.2). Its time series shows several anomalies, including a sharp peak during Carnival 2023 (*Temporal View* in Fig. 8). Node 553 also stands out in the *Correlation View*, while neighbors remain near the origin (Fig. 8B). The *Neighborhood View* highlights a few direct connections but many nearby crime-linked amenities — such as metro stations, bars, and pubs — that plausibly increase exposure and opportunity.

Peak weeks analysis. The peak of incidents spans two consecutive weeks: the week of Carnival and the week preceding it. The Carnival-week peak is predictably high, and the LLM correlates it with relevant contextual cues (amenities and nearby events), citing sources such as:

- “*The Pagu Block paraded on Tuesday (21) during Carnival at Praça da República in São Paulo.*”³
- “*Crowds with distracted people during their leisure time are a magnet for opportunistic actions... thefts and robberies increase during Carnival.*”⁴

The week-before-Carnival anomaly, however, may seem unexpected — especially for non-local users. Here, the LLM response is useful because it surfaces a plausible context not explicitly encoded in the dataset. For example, it identifies pre-Carnival events that may contribute to the early spike: “*The Ilú Obá de Min block opens the São Paulo carnival today (17) with a street opera. The parade will take place at night in the center of the capital’s streets.*”⁵

Takeaway. In cities like São Paulo, Carnival-related street events often start before the official dates. This type of contextual knowledge may be familiar to locals but is not universal, and it illustrates how grounded narrative explanations can help analysts interpret anomalies beyond purely statistical deviations.

³<https://www.alamy.com/image527174463.html>

⁴<https://movi-news.com/security/risks-during-the-carnival/07/02/2023/>

⁵<https://agenciabrasil.ebc.com.br/geral/noticia/2023-02/ilu-oba-de-min-abre-carnaval-paulistano-com-opera-de-rua>

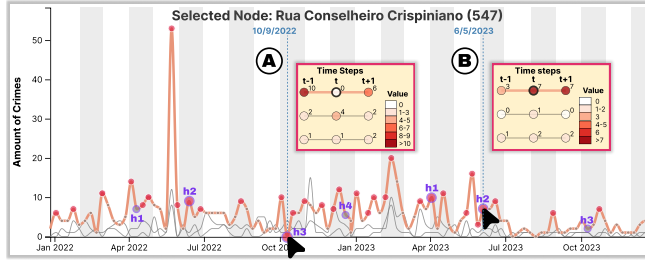


Fig. 9. Holiday-related anomalies. (A) Negative anomaly (dip). (B) Positive anomaly (peak). Marked Catholic holidays are (h1) Holy Week, (h2) Corpus Christi, (h3) Our Lady of Aparecida, and (h4) Christmas. © 2025 IEEE. Reprinted, with permission, from Heredia *et al.*, *Proc. SIBGRAPI* 2025 [17].

7.2.2 Religious holiday anomalies. In our initial exploration, most detected anomalies were *positive*, i.e., sudden increases in incident counts — expected, as low-incident streets make any spike noticeable. However, a pronounced *decrease* can be equally anomalous, especially for locations with a historically high baseline.

To examine this case, we focus on the upper-right quadrant of the *Correlation View* and select node 547 (Rua Conselheiro Crispiniano; highlighted in Fig. 4). Its time series contains several anomalies, including a marked dip (Fig. 9). The decline coincides with the holiday of *Our Lady of Aparecida*. Because a Catholic church is situated nearby, one plausible interpretation is that religious observance reduced routine street activity and, thus, opportunities for theft. However, does every major Catholic holiday produce a similar effect?

Comparative holiday analysis. Fig. 9 also marks three other Catholic holidays (labels h1–h4). To reduce tourism-related confounds (stronger in h1 and h4), we compare h3 in 2022 with h2 in 2023:

- **h3—Our Lady of Aparecida 2022** (Fig. 9A): the node recorded *zero* incidents during the holiday week.
- **h2—Corpus Christi 2023** (Fig. 9B): incidents spiked during the holiday week and remained elevated thereafter.

Node 547 is situated in a high-traffic corridor lined with amenities, making it structurally exposed to theft. Nevertheless, the LLM generated distinct but plausible explanations:

h3: “Likely due to religious observances associated with the Feast of Our Lady of Aparecida, leading to increased community presence and vigilance.” Consistent with “Our Lady of Aparecida coincides with Children’s Day in Brazil, leading to gatherings and a reflective, family-oriented atmosphere.”⁶

h2: “Heightened social activities and gatherings during Corpus Christi created opportunities for crime, both on the node itself and in nearby Praça Ramos de Azevedo.” Supported by “Guide to events and activities in São Paulo over the Corpus Christi long weekend.”⁷

Takeaway. Religious holidays do not uniformly suppress crime — even near churches. Whether incident counts rise or fall depends on the form of the observance: contemplative events may reduce routine street activity, whereas festive or touristic gatherings can generate the opposite effect.

7.3 Scenario 2: Manhattan Crashes

Exploring Manhattan, we observed that crash counts decrease over the years, with a pronounced drop in early 2020 (Fig. 10B). A plausible explanation is the COVID-19 pandemic and the preventative

⁶<https://www.mintageworld.com/media/detail/10602-the-day-of-our-lady-of-aparecida/>

⁷<https://www.guiadasemana.com.br/na-cidade/galeria/o-que-fazer-sao-paulo-feriado-corpus-christi-2023>

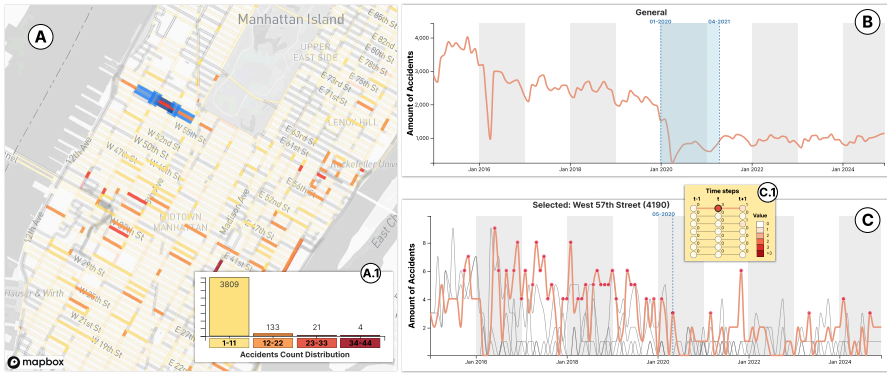


Fig. 10. COVID-19 in Manhattan (2020–early 2021). (A) Spatial View filtered to the selected time period (A.1 shows the legend). (B) Network-level Temporal View showing the overall drop after the pandemic onset. (C) Example street (W 57th St.) with historically higher crash counts, including a peak during the emergency.

measures adopted worldwide, such as travel restrictions, quarantines, and the promotion of remote work. When exploring street segments in the interface, we also found fewer detected anomalies after 2020, consistent with reduced congestion and lower baseline traffic volume. This suggests that unusual spikes may be more strongly associated with specific events or disruptions that concentrate incidents in a short period and area.

We select a street segment with many detected anomalies on *West 57th Street* (Fig. 10A). During the first months of the pandemic, we observed a sustained decrease, with one prominent peak in May 2020 (Fig. 10C.1). This segment does not stand out for nearby amenities; it includes a small number of hotels and restaurants. As a result, local PoI cues alone provide limited evidence for explaining the anomalous increase in crashes that month.

Peak analysis. By selecting this anomaly, the LLM proposes a hypothesis relating the peak to reopening announcements, supported by references and contextual statements such as:

- “Governor Cuomo announced that several regions, including New York City, began reopening starting in May and June 2020.”^{8,9}
- “West 57th Street is connected to other streets with varying accident rates, indicating possible changes in traffic flow and a higher risk of accidents than other cases.”

This response provides a checkable hypothesis that can be further examined by comparing nearby segments, alternative time windows, and additional sources.

Takeaway. Uncommon events, such as pandemic-related announcements, can provide critical temporal context for interpreting spikes in crash data. During periods with an overwhelming volume of reports and official updates, grounded narrative explanations help analysts identify plausible drivers and prioritize validation efforts.

8 Discussion and Limitations

We summarize the main findings, position the proposed framework within interactive visual analytics and narrative-centered AI, and discuss limitations and future directions for reliability, evaluation, deployment, and generalization.

⁸<https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-announces-five-regions-will-begin-reopening-today>

⁹<https://www.governor.ny.gov/news/governor-cuomo-announces-new-york-city-enter-phase-1-reopening-june-8-and-five-regions-enter>

8.1 Discussion

Interpretability and sensemaking. Our results suggest that LLM-generated narratives, when grounded in local ST context and coordinated visual cues, can support hypothesis formation during anomaly exploration. Rather than providing final explanations, the system acts as a sensemaking aid, helping users connect anomalous patterns with plausible external factors such as holidays, mass events, or spatial amenity distributions. This framing aligns the system more closely with exploratory visual analytics and interactive storytelling than with automated causal inference.

Role of structured prompting. A key design choice is the use of structured prompts that explicitly encode spatial, temporal, and historical context. To assess the contribution of this structure, we conducted a small-scale prompt ablation study in which key components were selectively removed and evaluated using an LLM-as-a-judge protocol (Sec. 6). The results suggest that richer structured context tends to be preferred over incomplete variants, which is consistent with its role as a stabilizing influence on narrative generation. Nevertheless, this structure can constrain expressiveness, highlighting a trade-off between output diversity and reliability.

Expert feedback and practical relevance. We gathered preliminary feedback through recorded sessions with two domain experts in the study of urban violence in São Paulo, part of an academic group that collaborates with the municipality; we presented the tool and walked through the usage scenarios while the experts raised questions throughout and at the end. Participants consented to the recording and to the use of anonymized quotes. Both experts indicated that the system’s main value lies in accelerating early-stage analysis, emphasizing the usefulness of rapid contextualization when screening anomalies in operational settings—as one of them put it, *“This would facilitate the discarding of some ideas, or also the exploration and validation of the most creative ones.”* At the same time, they noted that richer auxiliary data sources and more robust anomaly detectors would be required before deployment in security offices. We stress that this feedback is preliminary and qualitative rather than a controlled user study, indicating practical interest rather than measured user benefit.

8.2 Limitations and Future Work

While the proposed framework demonstrates the feasibility of grounded narrative explanations for ST anomalies, several limitations remain. We discuss these limitations and outline directions to strengthen reliability, improve empirical validation, and broaden applicability.

LLM variability and grounding limits. A key limitation is the variability of LLM-generated explanations. Even with structured prompts and context, explanations may differ in focus or include weakly supported claims. Our design mitigates these risks by constraining generation to localized ST context and requiring structured outputs, making narratives inspectable and actionable. While external web search sometimes provided useful context, it often lacked local or time-specific information. Future work can improve explanation stability through stronger grounding in structured contextual data. This includes tighter retrieval pipelines based on curated event repositories, city calendars, or vetted local news sources, reducing reliance on broad web search. Additional improvements could require each claim to be linked to explicit evidence (e.g., amenity cue or neighbor statistic), or enforce uncertainty statements when evidence is insufficient. Finally, interfaces that compare competing hypotheses (e.g., side-by-side narratives, hypothesis ranking, or contradiction highlighting) may help users detect instability and converge on better-supported explanations.

Limited user evaluation. A second limitation is the lack of a formal user study. While the interface follows established visual analytics principles and qualitative scenarios suggest practical value, we have not yet measured its impact on user understanding, trust, or decision-making in controlled settings. A systematic evaluation with both domain experts and non-experts is an

important next step. We envision a controlled study in which participants perform anomaly-analysis tasks with GROUNDSTORY and with a baseline condition lacking grounded narratives, combining task-performance measures with think-aloud protocols and standardized trust and usability questionnaires. Beyond usability and task-completion time, future studies should examine narrative-specific outcomes, such as perceived grounding, how users validate model-proposed hypotheses, and whether the system helps users generate and discard hypotheses more efficiently than baseline workflows. We also see value in longitudinal studies in which analysts revisit the same areas over time, enabling evaluation of whether narrative explanations improve recall, reduce confirmation bias, or support knowledge transfer across users.

Attribution and “ground truth” challenges. Assessing explanation correctness remains challenging, as real-world anomalies rarely have a single verifiable cause. This is particularly true in urban contexts where multiple exogenous factors can co-occur, and available evidence may be incomplete. Rather than identifying a unique cause, our goal is to help analysts formulate and validate plausible hypotheses grounded in observable context. Alternative evaluation approaches, such as counterfactual stress tests or controlled event injection studies, may also provide a more principled way to assess whether the narrative pipeline recovers known drivers.

Toward real-world deployment. Deploying GROUNDSTORY in operational urban analytics requires addressing several practical challenges. The pipeline depends on offline preprocessing (graph construction, anomaly detection, and context retrieval), which must be revalidated for new cities and data sources due to varying detector behavior and contextual signals. As noted by the experts, richer auxiliary data and analyst training are also needed. Since generated narratives are hypotheses rather than verified facts, human oversight remains essential for validation and uncertainty assessment [32]. Operational use also requires appropriate governance, access control, and auditing [18]. A pilot deployment co-designed with analysts is a promising next step.

Generalization beyond urban data. Beyond urban analytics, the approach can be applied to other domains such as environmental monitoring, epidemiology, and neural data analysis, where interpretable ST analysis is equally important. Future work could extend the framework to forecast-aware explanations that communicate predicted trends, uncertainty, and expected seasonal effects alongside observed anomalies.

9 Conclusion

We presented GROUNDSTORY, an interactive visual analytics framework that reframes urban ST anomaly explanation as graph-grounded, interactive storytelling, combining coordinated visualizations with LLM-generated narratives constrained by localized ST evidence. Across incident datasets from São Paulo and New York City, GROUNDSTORY produced contextual, human-readable hypotheses that users can inspect and validate through linked spatial and temporal views. Our exploratory prompt analysis further suggests that structured, graph-derived context tends to improve the relevance and grounding of generated explanations. Overall, these results indicate that narrative generation, when constrained by localized ST evidence and coupled with interactive exploration, can lower the barrier to hypothesis formation in complex urban data while preserving analyst control. Future work will focus on improving explanation stability, expanding evaluation with domain experts and non-experts, and extending the approach to additional ST domains.

Acknowledgments

This work was supported by FAPESP (#2021/07012-0, #2023/04868-7), FAPERJ (#E-26/210.585/2025), CNPq (#311144/2022-5), CAPES (#88887.184850/2025-00) Finance Code 001, and FGV. The opinions, hypotheses, conclusions, and recommendations expressed in this material are the responsibility of the authors and do not necessarily reflect the views of FAPESP, FAPERJ, CNPq, CAPES, and FGV.

References

- [1] Samuel Ackerman, Gal Amram, Ora Nova Fandina, Eitan Farchi, Shmulik Froimovich, Raviv Gal, Wesam Ibraheem, and Avi Ziv. 2025. LaajMeter: A Framework for Laaj Evaluation. *arXiv e-prints* (2025). doi:10.48550/arXiv.2508.10161
- [2] Nurnabi Meherul Alam, Sabyasachi Mitra, Surendra Kumar Pandey, Chayna Jana, Mrinmoy Ray, Sourav Ghosh, Sonali Paul Mazumdar, S. Vishnu Shankar, Ritesh Saha, and Gouranga Kar. 2024. Enhanced Spatio-Temporal Modeling for Rainfall Forecasting: A High-Resolution Grid Analysis. *Water* 16, 13 (2024), 1891. doi:10.3390/w16131891
- [3] Anthropic. 2024. Use XML tags in prompts. <https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/use-xml-tags>. Accessed: 2025-08-26.
- [4] Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A Survey on RAG with LLMs. *Procedia Computer Science* 246 (2024), 3781–3790. doi:10.1016/j.procs.2024.09.178
- [5] Geoff Boeing. 2025. Modeling and Analyzing Urban Networks and Amenities With OSMnx. *Geographical Analysis* 57, 4 (2025), 567–577. doi:10.1111/gean.70009
- [6] Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep learning for anomaly detection: A survey. *arXiv e-prints* (2019). doi:10.48550/arXiv.1901.03407
- [7] M. Crovella and E. Kolaczyk. 2003. Graph wavelets for spatial traffic analysis. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM-03, Vol. 3)*. 1848–1857. doi:10.1109/infcom.2003.1209207
- [8] Leyan Deng, Defu Lian, Zhenya Huang, and Enhong Chen. 2022. Graph Convolutional Adversarial Networks for Spatiotemporal Anomaly Detection. *IEEE Trans. Neural Networks and Learning Systems* 33, 6 (2022), 2416–2428. doi:10.1109/tnnls.2021.3136171
- [9] Zikun Deng, Di Weng, Xiao Xie, Jie Bao, Yu Zheng, Mingliang Xu, Wei Chen, and Yingcai Wu. 2022. Compass: Towards Better Causal Analysis of Urban Time Series. *IEEE Trans. Visualization and Computer Graphics* 28, 1 (2022), 1051–1061. doi:10.1109/tvcg.2021.3114875
- [10] Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a Graph: Encoding Graphs for Large Language Models. In *International Conference on Learning Representations*, Vol. 2024. 43909–43934.
- [11] Yuanning Feng, Sinan Wang, Zhengxiang Cheng, Yao Wan, and Dongping Chen. 2025. Are We on the Right Way to Assessing LLM-as-a-Judge? *arXiv e-prints* (2025). doi:10.48550/arXiv.2512.16041
- [12] Germain Garcia, Jaqueline Silveira, Jorge Poco, Afonso Paiva, Marcelo Batista Nery, Claudio T. Silva, Sergio Adorno, and Luis Gustavo Nonato. 2021. CrimAnalyzer: Understanding Crime Patterns in São Paulo. *IEEE Trans. Visualization and Computer Graphics* 27, 4 (2021), 2313–2328. doi:10.1109/tvcg.2019.2947515
- [13] Ali Hamdi, Khaled Shaban, Abdelkarim Erradi, Amr Mohamed, Shakila Khan Rumi, and Flora D. Salim. 2022. Spatiotemporal data mining: A survey on challenges and open problems. *Artificial Intelligence Review* 55, 2 (2022), 1441–1488. doi:10.1007/s10462-021-09994-y
- [14] David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. 2011. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* 30, 2 (2011), 129–150. doi:10.1016/j.acha.2010.04.005
- [15] Waqar Hassan, Marvin Mendes Cabral, Thiago Rodrigo Ramos, Antonio Castelo Filho, and Luis Gustavo Nonato. 2024. Modeling and predicting crimes in the city of São Paulo using graph neural networks. In *Brazilian Conference on Intelligent Systems*. Springer, 372–386. doi:10.1007/978-3-031-79035-5_26
- [16] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann Lecun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. In *Advances in Neural Information Processing Systems* 37. 132876–132907. doi:10.52202/079017-4224
- [17] Juanpablo Heredia, Leighton Estrada-Rayme, Jeremy Matos-Cangalaya, and Jorge Poco. 2025. Interactive Exploration and Explanation of Spatio-Temporal Anomalies with Graph-LLM Integration. In *2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. Institute of Electrical and Electronics Engineers (IEEE), 1–6. doi:10.1109/SIBGRAPI67909.2025.11223398
- [18] John S. Hollywood, Dulani Woods, Samuel Peterson, Michael J. D. Vermeer, and Brian A. Jackson. 2024. *Using Data Governance and Data Management in Law Enforcement: Building a Research Agenda That Includes Strategy, Implementation, and Needs for Innovation*. Research Report RR-A108-22. RAND Corporation, Santa Monica, CA. doi:10.7249/rra108-22
- [19] Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2025. GRAG: Graph Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 4145–4157. doi:10.18653/v1/2025.findings-naacl.232
- [20] Yue Jiang, Qin Chao, Yile Chen, Xiucheng Li, Shuai Liu, and Gao Cong. 2024. UrbanLLM: Autonomous Urban Activity Planning and Management with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 1810–1825. doi:10.18653/v1/2024.findings-emnlp.98
- [21] Ming Jin, Yaxuan Kong, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiao-Li Li, Vincent Tseng, Yu Zheng, Lei Chen, Hui Xiong, Shirui Pan, and Qingsong Wen. 2026. Large Models for Time

- Series and Spatio-Temporal Data: A Survey and Outlook. *Comput. Surveys* 58, 14 (2026), 1–38. doi:10.1145/3821637
- [22] Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. 2025. How to Correctly Report LLM-as-a-Judge Evaluations. *arXiv e-prints* (2025). doi:10.48550/arXiv.2511.21140
- [23] Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful?. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 9210–9232. doi:10.18653/v1/2023.findings-emnlp.618
- [24] Zhonghang Li, Lianhao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. UrbanGPT: Spatio-Temporal Large Language Models. In *Proc. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 5351–5362. doi:10.1145/3637528.3671578
- [25] Leonardo Padilla, Bernado Lagos-Álvarez, Jorge Mateu, and Emilio Porcu. 2020. Space-time autoregressive estimation and prediction with missing data based on Kalman filtering. *Environmetrics* 31, 7 (2020). doi:10.1002/env.2627
- [26] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohu Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2025. Graph Retrieval-Augmented Generation: A Survey. *ACM Trans. Inf. Syst.* 44, 2 (2025), 52 pages. doi:10.1145/3777378
- [27] Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. 2024. Let Your Graph Do the Talking: Encoding Structured Data for LLMs. *arXiv e-prints* (2024). doi:10.48550/arXiv.2402.05862
- [28] Santosha Rathod, Bishal Gurung, KN Singh, and Mrinmoy Ray. 2018. An improved space-time autoregressive moving average (STARMA) model for modelling and forecasting of spatio-temporal time-series data. *Journal of the Indian Society of Agricultural Statistics* 72, 3 (2018), 239–253.
- [29] Zhiyu Ren, Xiaojie Li, Jing Peng, Ken Chen, Qushan Tan, Xi Wu, and Canghong Shi. 2024. Graph autoencoder with mirror temporal convolutional networks for traffic anomaly detection. *Scientific Reports* 14, 1 (2024). doi:10.1038/s41598-024-51374-3
- [30] Elio Rodriguez, Nivan Ferreira, and Jorge Poco. 2022. JamVis: exploration and visualization of traffic jams. *The European Physical Journal Special Topics* 231, 9 (2022), 1673–1687. doi:10.1140/epjs/s11734-021-00424-2
- [31] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. 2021. A Unifying Review of Deep and Shallow Anomaly Detection. *Proc. IEEE* 109, 5 (2021), 756–795. doi:10.1109/jproc.2021.3052449
- [32] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. 2016. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Trans. Visualization and Computer Graphics* 22, 1 (2016), 240–249. doi:10.1109/tvcg.2015.2467591
- [33] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv e-prints* (2017). doi:10.48550/arXiv.1708.08296
- [34] Seyyid Emre Sofuoglu and Selin Aviyente. 2022. GLOSS: Tensor-based anomaly detection in spatiotemporal urban traffic data. *Signal Processing* 192 (2022), 108370. doi:10.1016/j.sigpro.2021.108370
- [35] Zhiwen Tian, Ming Zhuo, Leyuan Liu, Junyi Chen, and Shijie Zhou. 2023. Anomaly detection using spatial and temporal information in multivariate time series. *Scientific Reports* 13, 1 (2023). doi:10.1038/s41598-023-31193-8
- [36] Leo Tisljarić, Sofia Fernandes, Tonči Carić, and João Gama. 2020. Spatiotemporal Traffic Anomaly Detection on Urban Road Network Using Tensor Decomposition Method. In *Discovery Science*. Springer International Publishing, 674–688. doi:10.1007/978-3-030-61527-7_44
- [37] Francesco Verdoja and Marco Grangetto. 2020. Graph Laplacian for image anomaly detection. *Machine Vision and Applications* 31, 1-2 (2020). doi:10.1007/s00138-020-01059-4
- [38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 24824–24837. doi:10.52202/068431-1800
- [39] Huichu Zhang, Yu Zheng, and Yong Yu. 2018. Detecting Urban Anomalies Using Multiple Spatio-Temporal Data Sources. *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–18. doi:10.1145/3191786
- [40] Mingyang Zhang, Tong Li, Yue Yu, Yong Li, Pan Hui, and Yu Zheng. 2022. Urban Anomaly Analytics: Description, Detection, and Prediction. *IEEE Trans. Big Data* 8, 3 (2022), 809–826. doi:10.1109/tbdata.2020.2991008
- [41] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2026. A comprehensive survey on automatic text summarization with exploration of LLM-based methods. *Neurocomputing* 663 (2026), 131928. doi:10.1016/j.neucom.2025.131928

Received 16 February 2026; revised 27 June 2026; accepted 10 September 2026