

CaST: Causal Discovery via Spatio-Temporal Graphs in Disaster Tweets

Hieu Minh Duong and Long Nguyen
J.B. Speed School of Engineering
University of Louisville
Louisville, KY, USA
{hieu.duong, l.nguyen}@louisville.edu

Eugene Levin and Todd Gary
School of Applied Computational Sciences
Meharry Medical College
Nashville, TN, USA
{elevin, tpgary}@mmc.edu

Abstract—Understanding causality between real-world events from social media is essential for situational awareness, yet existing causal discovery methods often overlook the interplay between semantic, spatial, and temporal contexts. We propose CaST: Causal Discovery via Spatio-Temporal Graphs, a unified framework for causal discovery in disaster domain that integrates semantic similarity and spatio-temporal proximity using Large Language Models (LLMs) pretrained on disaster datasets. CaST constructs an event graph for each window of tweets. Each event extracted from tweets is represented as a node embedding enriched with its contextual semantics, geographic coordinates, and temporal features. These event nodes are then connected to form a spatio-temporal event graph, which is processed using a multi-head Graph Attention Network (GAT) [1] to learn directed causal relationships. We construct an in-house dataset of approximately 167K disaster-related tweets collected during Hurricane Harvey and annotated following the MAVEN-ERE schema. Experimental results show that CaST achieves superior performance over both traditional and state-of-the-art methods. Ablation studies further confirm that incorporating spatial and temporal signals substantially improves both recall and stability during training. Overall, CaST demonstrates that integrating spatio-temporal reasoning into event graphs enables more robust and interpretable causal discovery in disaster-related social media text.

Keywords—Computing Methodologies, Artificial Intelligence, Natural Language Processing, Information Extraction

I. INTRODUCTION

Causal discovery aims to identify cause–effect relationships among events from observational data, offering deeper insights into how complex phenomena unfold. In disaster scenarios, understanding causality across space and time is critical for improving situational awareness, forecasting cascading impacts, and supporting decision-making for emergency response. Social media platforms such as Twitter have become valuable data sources in this context, as they provide real-time, fine-grained signals of unfolding disaster events, ranging from infrastructure damage to community-level responses. These data streams often reflect complex, interdependent phenomena that evolve rapidly across locations and timescales. Figure 1 illustrates how disaster-related events unfold across time and space in social media text. In this example, one tweet describes heavy rain leading to flooding and traffic jams, while another

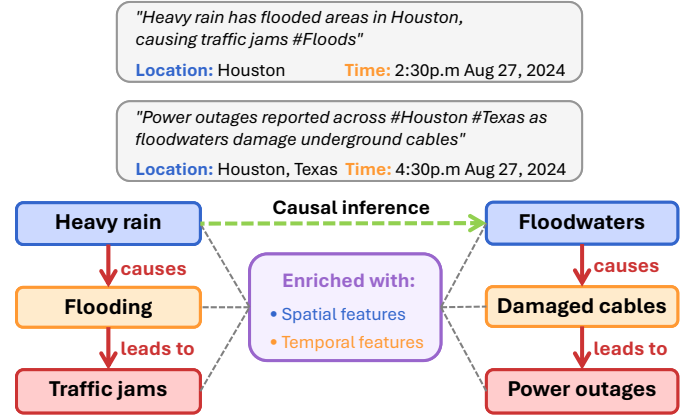


Fig. 1. Example of causal relationships in disaster tweets with spatio-temporal context. Two tweets from the same location (Houston) and close temporal proximity (2 hours apart) contain causally related events. The green dashed arrow illustrates the causal inference between "Heavy rain" and "Floodwaters" across tweets, enabled by shared spatio-temporal features. Gray dashed lines show how spatial and temporal context enriches event representations to improve causal discovery.

later tweet reports floodwaters causing damaged cables and power outages in the same region. Although these tweets are posted separately, they represent a continuous chain of disaster effects occurring within a short temporal window and close spatial proximity. Such examples highlight the need for spatio-temporal reasoning in causal discovery: existing text-based approaches often capture only intra-tweet relations, overlooking cross-tweet causal propagation that frequently occurs in real-world disaster scenarios.

However, inferring causality from such unstructured, dynamic, and context-rich text remains a challenging task. Traditional causal discovery approaches, including statistical and graph-based models, often fail to capture the intricate spatial and temporal dependencies that characterize real-world disaster information. Recent advancements in Large Language Models (LLMs) such as BERT and GPT have shown exceptional capabilities in semantic understanding, enabling extraction of event representations from unstructured text. Yet, most existing LLM-based methods for causal inference focus solely on textual relationships and ignore the spatio-temporal

interdependencies among events, limiting their ability to capture the real-world propagation patterns typical of disasters. In particular, events such as *floods causing power outages*, *aftershocks triggering new collapses*, or *road closures leading to rescue delays* require both temporal reasoning (when events happen) and spatial awareness (where they occur) to be causally understood. Integrating these dimensions is therefore essential to identify how disaster effects spread and interact.

Despite recent advances, current approaches often fall short in integrating both spatio-temporal reasoning within a unified causal discovery pipeline. Many graph-based models overlook temporal dependencies or treat cross-document causality as a secondary task, while prompt-based LLM methods, though flexible, lack structural consistency and are prone to hallucination. Likewise, methods that inject external knowledge or event semantics often do so statically, without modeling how events evolve and interact across time and space—an essential aspect in domains like disaster response, where causal chains unfold dynamically across locations and hours. In large-scale events such as hurricanes, floods, or wildfires, understanding where and when cascading effects occur is vital for situational awareness and coordinated response. However, few existing methods explicitly capture such spatio-temporal propagation patterns in social media streams. To address these challenges, we aim to develop a unified framework that integrates semantic, spatial, and temporal knowledge into a single graph-learning paradigm. In this framework, each event extracted from social media text is represented as a node enriched not only with linguistic features from LLM embeddings but also with contextual signals derived from its spatial location and temporal occurrence. The event graph is then constructed and expanded along two dimensions: (i) *Spatial neighbors*, capturing proximity in location, and (ii) *Temporal neighbors*, capturing alignment in time. This enriched graph is subsequently processed using a multi-head Graph Attention Network (GAT), which enables dynamic attention across spatial and temporal relations. Through this design, CaST bridges the gap between text-based causality and real-world disaster dynamics, allowing the model to learn context-aware causal representations that better reflect event evolution across time and space. Our contributions are summarized as follows:

- We propose CaST (Causal Discovery via Spatio-Temporal Graphs), a novel framework for causal discovery that unifies spatial and temporal relationships into a spatio-temporal event graph derived from disaster-related social media data. Unlike prior methods that rely purely on textual cues, CaST explicitly captures how disaster-related events unfold across time and locations.
- We construct a large-scale dataset of approximately 167K disaster-related tweets, annotated following the MAVEN-ERE schema with extensions for temporal, spatial, and causal information. This dataset enables fine-grained evaluation of causal inference in real-world, noisy social media contexts.
- We assess CaST against eight strong baselines, ranging

from traditional approaches to graph-based, neural, and prompt-based models. Results show that CaST achieves the best balance between precision and recall, outperforming prior methods in both causal link identification and generalization stability. Ablation studies further confirm the importance of spatial and temporal representations in enhancing causal inference.

II. RELATED WORK

A. Early Approaches for Causal Discovery

Early methods identified causal relations in text using patterns and linguistic cues. Khoo et al. [2] applied manual templates such as “X leads to Y” to newspaper articles to extract cause–effect pairs. Supervised feature-based classifiers were later introduced: Blanco et al. [3] trained a classifier on lexical and syntactic features to detect causal relations. Researchers also began creating evaluation datasets; the SemEval-2010 Task 8 included Cause–Effect as one of several semantic relation classes. Girju [4] developed an early question–answering system for causal QA, using semantic parsing and WordNet knowledge to detect causality in text. With the rise of machine learning and larger datasets, causal extraction methods shifted toward neural networks and external knowledge integration. Cao et al. [5] proposed a knowledge-enriched model (LSIN) that incorporates external graphs for event causality identification. Their Latent Structure Induction Networks use two modules: a Descriptive Graph Induction, which pulls descriptive attributes of events from knowledge bases, and a Relational Graph Induction, which learns latent multi-hop connections between events.

Data augmentation and transfer learning were also explored. Cao et al. and Liu et al. generated synthetic causal sentences or mined additional training examples from external corpora. Cao et al. [5] leveraged patterns from lexical knowledge bases (WordNet, VerbNet) to create paraphrases of causal statements and used contrastive learning to teach a model invariant features of causality. Neural models were further extended to jointly model related tasks. Shen et al. [6] introduced a prompt-based multi-task learning approach for event causality identification. They crafted auxiliary prompt tasks that a pretrained language model must answer about the presence of causal cues and plausible causal links, drawing on latent knowledge of causation in the language model. Zuo et al. [7] present LearnDA, a dual-learning augmentation framework for causal sentence generation. Transfer-learning strategies are examined by Anuyah et al. [8], who show that combining data from multiple domains improves causal relation extraction performance. More recently, Chun et al. [9] leverage pretrained language models (e.g., RoBERTa, T5) for data augmentation in event causality tasks, demonstrating benefits even under label imbalance. These studies highlight evolving trends: beyond vanilla supervised models, augmentation and cross-domain transfer hold promise for improving causality extraction in low-resource or implicit-causal settings.

B. Graph-Based and Joint Reasoning Methods

Recent state-of-the-art approaches emphasize structured reasoning using graphs to represent events and their relations. Instead of treating each candidate event pair independently, these methods model global structures where multiple causal links interact and support logical consistency across a document. Gao et al. [10] introduced one of the earliest structured approaches, building an Integer Linear Programming (ILP) framework to enforce global constraints on causal relations at the document level. Their system identified key central events and applied constraints such as transitivity and consistency with discourse cues to capture coherent causal chains within long narratives.

Following this, neural graph-based models were developed to integrate structured reasoning directly into the model architecture. Graph Neural Networks (GNNs) became the primary tool for this purpose. Models such as those by Phu and Nguyen [11] and Fan et al. [12] represented each document as an event graph, where nodes correspond to events and edges represent potential causal or semantic relations. Through message passing, information from surrounding events, coreference links, and discourse relations informed the prediction of causal links within and across sentences. Liu et al. [13] extended this idea by constructing directed causal graphs and iteratively refining them through an iterative Learning and Identifying Framework (iLIF), which first identified high-confidence causal links and then updated event representations through a graph encoder before repeating the process.

Building on these developments, Pu et al. [14] proposed a joint framework to extract events and causal relations simultaneously. Their model employed a heterogeneous relation graph that included event–event edges for causality, as well as event–argument and argument–argument edges for shared entities and semantic roles. Event types and words formed the graph nodes, and a GNN captured the interactions among event identity, arguments, and candidate causal connections. They also introduced a multi-channel label enhancement strategy to represent the cause and effect roles within this unified architecture.

To enrich the semantic structure of graph-based models, Hu et al. [15] incorporated Abstract Meaning Representation (AMR) to provide deeper semantic context for causal inference. Sentences were converted into AMR graphs encoding the conceptual and relational structure of events, and information from AMR subgraphs connecting event mentions was aggregated to construct a semantic reasoning graph. Their model, SemSIn, used a GNN to integrate event-centric structures with an LSTM-based path encoder for event-associated structures, allowing the representation of implicit relations through underlying semantic links.

C. Spatio-Temporal and Domain-Specific Causal Discovery

A special subclass of causal text mining focuses on integrating temporal and spatial information to understand when and where events occur. Mirza and Tonelli were among the first to jointly analyze temporal and causal relations in text [16].

They proposed a framework that links event causality with temporal ordering, using the principle that causes generally precede effects in time and employing temporal relations as features for causal inference.

Liu et al. (2023) [17] introduced PPAT (Progressive Graph Pairwise Attention Network), which refines event pair representations through progressive attention layers and graph-based message passing. PPAT models causal directionality more effectively than GCN-based methods by learning hierarchical attention from token-level to event-level features. Liu et al. (2025) [18] proposed a model that jointly learns event causality and temporality. Their framework uses a dual-channel neural network, where one channel identifies causal relations between events and the other detects temporal before/after relations. The model constructs an event causality graph in which nodes are weighted by temporal salience, allowing temporal signals to inform causal reasoning and providing a unified representation of event dependencies.

Spatio-temporal modeling has also been extended to domain-specific contexts such as disaster analysis and social media monitoring. Dong et al. [19] developed a real-time framework for extracting cascading disaster effects from Twitter streams. Their system continuously collects posts during disasters, applies co-word analysis to detect emerging topics, identifies potential causal chains, and enriches them by extracting geolocation references to visualize geographically distributed consequences. In another domain, Lenti et al. [20] investigated causal discovery in social movements related to climate activism. They constructed a probabilistic causal graph integrating Reddit discussion data with external event information to model the factors influencing protest participation. The framework captures multi-factor causal influences, representing relationships such as interactions with activists, media exposure, and engagement intensity.

D. Large Language Models for Causal Discovery

Large Language Models (LLMs) such as GPT-3.5 and GPT-4 have introduced a new direction for causal reasoning from text. Kiciman et al. [21] evaluated these models across several benchmarks, including pairwise causal discovery, counterfactual reasoning, and event causality tasks. Their study demonstrated that LLMs can infer causal relationships directly from textual descriptions, generate causal graphs, and provide natural language explanations of causal links, reflecting their internalized commonsense knowledge acquired through large-scale pretraining.

Subsequent research has explored combining LLMs with traditional causal inference algorithms. Long et al. (2023) demonstrate this by using GPT-generated causal statements as priors or constraints for Bayesian network learners as in NLP applications, LLMs have been employed to extract candidate causal graphs from corpora that can later be refined through statistical modeling or human feedback [22]. This hybrid approach leverages the linguistic and world knowledge capabilities of LLMs while connecting them with structured causal frameworks. Recent work has further extended this

paradigm by integrating structured representations directly into LLM-based reasoning. Shyalika et al. [23] incorporated Causal Event Graphs as contextual input to LLMs to enhance spatial-temporal event reasoning. In this setup, LLMs process both textual and graph-based representations, combining natural language understanding with explicit relational structures for more interpretable causal inference.

III. METHODOLOGY

A. Problem Formulation

According to Hume’s definition of causality [24], three fundamental conditions must hold: A causes B if (1) A precedes B in time, (2) A and B are contiguous in space and time, and (3) A and B co-occur or neither occurs. Building upon this philosophical grounding, our framework operationalizes causality through explicit temporal precedence, spatial proximity, and contextual co-occurrence among detected events in disaster-related tweets. We formalize causal discovery from disaster tweets as an edge classification problem on a spatio-temporal event graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes represent extracted events enriched with semantic, spatial, and temporal features, and edges denote candidate causal connections. The task is to predict whether a directed edge $e_i \rightarrow e_j$ exists between two events based on their contextual and structural representations.

B. Framework Overview

Figure 2 illustrates the overall architecture of the proposed CaST framework. The procedural steps are organized into three modules—Feature Extraction (Algorithm 1), Spatio-Temporal Graph Construction (Algorithm 2), and Causal Learning with GAT (Algorithm 3). CaST begins by extracting event triggers, arguments, and their spatial and temporal attributes from disaster-related tweets. These elements are encoded into heterogeneous node representations and serve as the basis for constructing a unified spatio-temporal event graph. The graph incorporates semantic, spatial, and temporal relationships that link events across tweets within each time window. Once the graph is built, the model applies a multi-head Graph Attention Network to propagate information across these relations and obtain context-aware event representations. A final classifier then predicts directional causal links between event pairs, allowing CaST to infer causal dynamics that are grounded in both linguistic content and spatio-temporal context.

Each tweet is processed to extract event triggers and their arguments, along with associated spatial and temporal cues. These elements form heterogeneous nodes: event nodes capture semantic content, while spatial and temporal nodes encode contextual information such as location and timestamp. Before graph learning, CaST combines these features into unified event representations that integrate semantic, spatial, and temporal signals. The spatio-temporal event graph contains nodes for events and contextual attributes, with edges representing semantic, spatial, and temporal relations. The Graph Attention Network then propagates and aligns information across these connections, enabling the model to learn dependencies that

reflect how disaster-related events unfold. Finally, a causal link classifier determines whether a directed causal relation exists between any event pair, producing a structured representation of causal chains across time and space.

C. Feature Extraction

Algorithm 1 outlines the process used to extract event triggers, arguments, and their associated spatial-temporal attributes from the raw tweet stream. The steps in Algorithm 1 are implemented as follows.

Algorithm 1 Feature Extraction

Input: Tweet set \mathcal{T}

Output: Extracted events and initial event features

- 1: **for** each tweet $t_i \in \mathcal{T}$ **do**
 - 2: Extract event triggers and arguments using NER and patterns.
 - 3: Obtain temporal features τ_i and spatial features s_i .
 - 4: Construct initial event representations from semantic, spatial, and temporal features.
 - 5: **end for**
 - 6: **return** extracted events and event representations
-

Given a tweet t_i containing a sequence of tokens $\{w_1, w_2, \dots, w_n\}$, each annotated event trigger e_j is encoded into a dense representation using the CrisisTransformer model [25]. The contextualized embedding \mathbf{v}_{e_j} is derived by combining the trigger token embedding and the [CLS] token embedding as:

$$\mathbf{v}_{e_j} = \alpha \cdot \mathbf{v}_{\text{trigger}} + (1 - \alpha) \cdot \mathbf{v}_{[\text{CLS}]}, \quad (1)$$

where $\alpha = 0.7$ is empirically selected to balance local and global context. Each tweet is also associated with metadata providing temporal and spatial cues.

Temporal features \mathbf{t}_i are normalized indicators derived from the posting timestamp:

$$\mathbf{t}_i = \left[\frac{h}{24}, \frac{d}{7}, \frac{m}{12}, \frac{\text{day}}{31} \right], \quad (2)$$

where h , d , m , and day represent the hour, weekday, month, and day components, respectively. Spatial features \mathbf{s}_i capture both coordinate-based and text-based location signals extracted from each tweet:

$$\mathbf{s}_i = [g_1, g_2, l_1, l_2, \hat{\phi}, \hat{\lambda}], \quad (3)$$

where g_1 and g_2 are binary indicators denoting the presence of geolocation metadata (from tweet coordinates or bounding boxes) and explicit textual location mentions, respectively. l_1 represents the normalized count of distinct location mentions extracted from the tweet text, while $(\hat{\phi}, \hat{\lambda})$ denote the latitude and longitude values normalized to $[-1, 1]$ by dividing by 90 and 180, respectively. When geolocation information is missing, coordinate values are set to zero. This formulation enables CaST to incorporate both explicit spatial references in text and implicit geographic metadata where available.

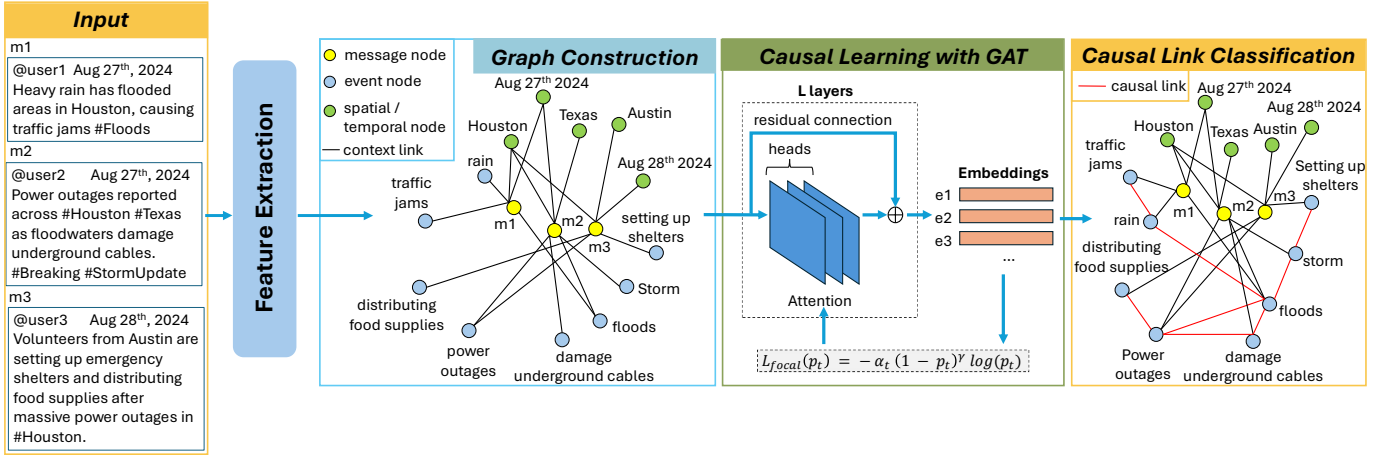


Fig. 2. Overview of the proposed CaST framework for causal discovery in disaster tweets. The Feature Extraction module extracts events along with triggers and contextual attributes (spatial and temporal attributes), encodes them as heterogeneous nodes. A spatio-temporal graph is then constructed where black edges denote contextual link (semantic, spatial, or temporal relation), and red edges indicate predicted causal links in the output. The Graph Attention Network propagates information across these edges to learn context-aware representations that enable accurate causal link classification.

The extracted event, spatial, and temporal embeddings correspond to distinct node types that form the heterogeneous graph. Before input to the GAT model, these representations are projected into a shared latent space and concatenated to form unified node features:

$$\mathbf{x}_{e_j} = [\mathbf{v}_{e_j} \parallel \mathbf{t}_i \parallel \mathbf{s}_i], \quad (4)$$

where \mathbf{v}_{e_j} , \mathbf{t}_i , and \mathbf{s}_i denote the event, temporal, and spatial feature vectors, respectively. This fusion ensures that each event node encodes both semantic and contextual information, enabling the GAT to learn richer spatio-temporal dependencies in subsequent causal graph learning.

D. Graph Construction

Algorithm 2 details how these extracted elements are organized into a unified spatio-temporal event graph within each temporal window.

Algorithm 2 Spatio-Temporal Graph Construction

Input: Processed tweets \mathcal{T} , time window w_t

Output: Set of spatio-temporal graphs $\{\mathcal{G}\}$

- 1: Partition tweets into chronological windows of size w_t .
- 2: **for** each window **do**
- 3: Initialize graph \mathcal{G} .
- 4: **for** each tweet t_i in the window **do**
- 5: Add event nodes to \mathcal{G} .
- 6: Link each event to its spatial and temporal nodes.
- 7: Add edges between events co-occurring in the same tweet.
- 8: **end for**
- 9: Add \mathcal{G} to the graph set.
- 10: **end for**
- 11: **return** $\{\mathcal{G}\}$

The Graph Construction module transforms event representations into a unified spatio-temporal graph that models

relationships among events, locations, and time references. Each tweet contributes: (1) *event nodes* representing detected event triggers and their arguments, (2) *spatial nodes* corresponding to extracted geographic entities such as cities or regions, and (3) *temporal nodes* denoting timestamps derived from tweet metadata or explicit temporal expressions. These nodes collectively encode the semantic, spatial, and temporal characteristics of disaster-related discourse.

To preserve chronological dependencies and manage scalability, CaST constructs individual graphs within a fixed temporal window of tweets (e.g., several hours). Tweets falling within the same window are grouped to form one graph, ensuring that events close in time are contextually connected while maintaining temporal ordering across graphs. This sliding-window approach captures local causal dynamics and supports incremental learning as new tweets arrive, reflecting the evolving nature of disaster events.

Edges are established to represent interdependencies among nodes. Contextual links connect event nodes that co-occur within the same tweet or exhibit high semantic similarity. Spatial links connect events occurring in geographically proximate regions, determined via geolocation metadata or named-entity alignment. Temporal links connect events according to their chronological order within each graph window, ensuring that causal direction respects temporal precedence. This design operationalizes Hume’s classical notions of causation—temporal priority and spatial contiguity—within a data-driven framework.

The resulting graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is heterogeneous, where \mathcal{V} includes event, spatial, and temporal nodes, and \mathcal{E} comprises semantic, spatial, and temporal relations. Each node $v_i \in \mathcal{V}$ is initialized with its corresponding embedding vector (semantic from the language model, spatial from location encoding, and temporal from timestamp encoding). These heterogeneous embeddings are aligned into a shared feature space and subse-

quently passed to the Graph Attention Network for relational reasoning and causal inference. This window-based graph construction allows CaST to model both local and evolving causal interactions among disaster events, effectively balancing temporal continuity and computational efficiency.

E. Causal Learning with Multi-head GAT model

Algorithm 3 summarizes the GAT-based causal learning procedure that updates node representations and predicts directional causal links.

Algorithm 3 Causal Learning and Link Classification

Input: Graphs $\{\mathcal{G}\}$, number of GAT layers L , number of mini-batches B , threshold δ

Output: Predicted causal links \mathcal{C}

```

1: for each mini-batch  $b = 1$  to  $B$  do
2:   for each GAT layer  $l = 1$  to  $L$  do
3:     Update node representations using attention-based
       message passing.
4:   end for
5:   Compute causal scores for candidate event pairs.
6:   Update model parameters using Focal Loss.
7: end for
8: Identify causal links using threshold  $\delta$ .
9: return predicted causal links  $\mathcal{C}$ 

```

We employ a two-layer Graph Attention Network (GAT) to model event interactions and capture causal dependencies across spatial, temporal, and contextual relations. Given node features \mathbf{x}_v , the GAT updates them through attention-weighted message passing as follows:

$$\mathbf{h}_v^{(l+1)} = \sigma \left(\sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_u^{(l)} \right), \quad (5)$$

where $\mathcal{N}(v)$ denotes the set of neighbors of node v , $\mathbf{h}_u^{(l)}$ is the node embedding at layer l (with $\mathbf{h}_u^{(0)} = \mathbf{x}_u$), $\mathbf{W}^{(l)}$ is a trainable transformation matrix, and $\alpha_{vu}^{(l)}$ is the learned attention coefficient. The nonlinearity $\sigma(\cdot)$ is the ELU activation function.

Attention coefficients are computed as:

$$\alpha_{vu}^{(l)} = \text{softmax}_u \left(\text{LeakyReLU}(\mathbf{a}^\top \mathbf{z}_{vu}^{(l)}) \right), \quad (6)$$

where $\mathbf{z}_{vu}^{(l)} = [\mathbf{W}^{(l)} \mathbf{h}_v^{(l)} \parallel \mathbf{W}^{(l)} \mathbf{h}_u^{(l)}]$ and \mathbf{a} is a trainable attention vector. Residual connections are applied between layers to stabilize training and maintain hierarchical consistency, as shown in Figure 2.

After the GAT layers, the final node embeddings of event pairs (v_i, v_j) are concatenated and passed into a fully connected classifier to estimate causal likelihood:

$$\hat{y}_{ij} = \text{softmax}(f_{\text{MLP}}([\mathbf{h}_{v_i} \parallel \mathbf{h}_{v_j}])), \quad (7)$$

where $f_{\text{MLP}}(\cdot)$ is a multi-layer perceptron that outputs class logits. The resulting distribution \hat{y}_{ij} indicates the predicted

probability of a causal or non-causal relation between the event pair.

The model is trained using the Focal Loss [26], which mitigates class imbalance by down-weighting easy negatives and emphasizing harder, misclassified causal links:

$$L_{\text{focal}}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad (8)$$

where p_t is the predicted probability corresponding to the ground-truth class, α_t is a balancing factor, and γ is the focusing parameter.

By jointly training with focal loss, CaST becomes more sensitive to minority causal cases and effectively reduces the impact of abundant non-causal pairs. This GAT-based causal learner corresponds to the ‘‘Training Multi-Head GAT Model’’ stage in Figure 2, integrating event-level semantics with spatial and temporal context for robust causal detection in disaster-related tweets.

F. Causal Link Classification

The final stage of CaST identifies directed causal relations between event pairs represented in the constructed graph. Each candidate pair is encoded by concatenating the learned event embeddings, which capture both contextual information and their structural roles within the graph. A fully connected classifier predicts whether a causal link exists between the two events, producing a directed causal graph that reflects underlying disaster propagation patterns across space and time.

IV. EXPERIMENTS

A. Dataset

We conduct experiments on an in-house dataset consisting of a total of $\sim 167\text{K}$ disaster-related tweets, each annotated with causal relations between events. The raw corpus originally contained several million tweets collected between August 26 and August 30, 2017, corresponding to the Hurricane Harvey disaster period. To construct a manageable and chronologically consistent dataset, we applied a sampling rate of 1:50, preserving temporal order. Relevant attributes such as tweet text, timestamp, and location information were filtered and stored in JSON format following the schema shown in Figure 3. The dataset was annotated following the MAVEN-ERE event schema [27], with minor adjustments for our task. Named Entity Recognition (NER) and linguistic pattern matching were first applied to extract candidate event triggers from tweets, which were then manually verified by two expert annotators. Annotators further examined each tweet to label directed causal links between events, specifying the cause–effect pairs. At the token level, a mask field marks causal roles using I-C (inside a cause span), I-E (inside an effect span), and O (outside any causal span). The final dataset includes: (1) event triggers and arguments, (2) temporal and geospatial metadata, (3) intra-tweet causal relations, and (4) span-level causal masks. This structure enables both event-level and edge-level causal discovery under the proposed CaST framework.


```

{
  "tweet_text": "<tweet text here>",
  "tokens": ["<token_1>", "<token_2>", "..."],
  "events": [
    {
      "id": "evt_XXX", "trigger": "<trigger_word>",
      "arguments": {
        "arg_1": "<argument_1>"
      }
    },
    {
      "id": "evt_YYY", "trigger": "<trigger_word>",
      "arguments": {
        "arg_1": "<argument_1>"
      }
    }
  ],
  "causal_relation": {
    "relation": <true_or_false>,
    "pairs": [{"CAUSE": "evt_XXX", "EFFECT": "evt_YYY"}]
  },
  "mask": ["O", "I-C", "I-E", "..."],
  "tweet_id": "<tweet_id>",
  "date_str": "<date_string>",
  "date_numeric": "<unix_timestamp>",
  "geolocation": "<location_string>",
  "bounding_box": "(<lat1>,<lon1>),(<lat2>,<lon2>),(<lat3>,<lon3>),(<lat4>,<lon4>)"
}

```

Fig. 3. JSON schema of the annotated disaster tweet dataset used in CaST. Each tweet includes the identification number, detected events, causal relations, spatial and temporal metadata, and token-level causal role masks.

Although the CaST framework is capable of modeling inter-tweet causal dependencies through its spatio-temporal graph design, our in-house dataset currently annotates only intra-tweet causal relations. Consequently, this study focuses on evaluating intra-tweet causal discovery performance, while cross-tweet (inter-tweet) causal reasoning is left for future work.

B. Experimental Setup

We trained the CaST framework on the disaster tweet dataset using an 80–10–10 split for training, validation, and testing, respectively. Each event graph was treated as a sample, and we employed mini-batch training with a batch size of 32 and a learning rate of 0.001. The model was trained for up to 100 epochs on an NVIDIA RTX 5000 Ada Generation GPU with 16 GB of memory. To address class imbalance between causal and non-causal pairs, we computed class weights based on their frequency ratio in the training data. The input feature dimension corresponds to the concatenated spatio-temporal embedding of each event node. The GNN backbone uses a hidden dimension of 256, 16 attention heads, and a dropout rate of 0.1. Model parameters were optimized using Adam, and training was stopped early if the validation loss did not improve for 10 consecutive epochs. The model achieving the lowest validation loss was selected for final evaluation on the held-out test set.

C. Baselines

We compare CaST against a range of representative baselines spanning classical machine learning, neural architectures,

transformer models, and graph-based causal discovery methods. To ensure a fair comparison, all baselines operate on the same event-pair representation. For each pair of events (e_i, e_j) , we extract their trigger words and construct an input sequence of the form:

$\text{trigger}_i [\text{SEP}] \text{trigger}_j [\text{SEP}] \text{tweet_text}.$

The CrisisTransformer embeddings of each event trigger are used whenever semantic embeddings are required. Spatial and temporal features are excluded for the baseline models to isolate the contribution of spatio-temporal reasoning in CaST.

- **Additive Noise Model (ANM)** [28]. A classical statistical causal discovery method that assumes the effect is a nonlinear function of the cause plus independent noise.
- **Random Forest (RF)**. We train a Random Forest classifier using the concatenated CrisisTransformer embeddings of (e_i, e_j) as features. The model uses 100 trees, a maximum depth of 10, `class_weight=balanced`, and Gini impurity. These settings follow standard practice and were validated on the development set.
- **Support Vector Machine (SVM)**. The SVM baseline uses an RBF kernel with $C = 1.0$ and $\gamma = 0.01$. The classifier takes the same concatenated event embeddings as input. The model is trained on the same 70–15–15 split as CaST.
- **BiLSTM + Attention** [29]. A bidirectional LSTM with two layers and hidden size 256 is used to encode the token sequence. An attention layer aggregates hidden states into an event-pair representation, followed by a 2-layer MLP classifier. We train the model for up to 15 epochs with learning rate 10^{-3} and Adam optimizer, selecting the best checkpoint based on validation F1.
- **BERT** [30]. We fine-tune `bert-base-uncased` for causal classification. The input sequence is tokenized as described above, and the model pools three signals: the [CLS] embedding, the mean embedding of event_i, and the mean embedding of event_j. These three vectors are concatenated and passed to a feed-forward classifier. BERT is trained with AdamW, learning rate 2×10^{-5} , batch size 16, and early stopping based on validation F1.
- **Document-level Event Causality Identification (DECI)** [11]. A graph-based neural architecture that applies graph convolutional layers over document-level event graphs. We follow the hyperparameter settings reported in the original paper: two GCN layers with hidden size 256 and learning rate 10^{-3} .
- **Progressive Graph Pairwise Attention Network (PPAT)** [17]. A neural architecture that refines event-pair representations through progressive attention layers. We use the authors' implementation with default hyperparameters, including hierarchical attention and pairwise refinement across layers.
- **DAPrompt** [31]. A prompt-based approach that fine-tunes a pretrained language model using deterministic assumption prompts. We use the recommended configu-

ration from the original paper with learning rate 3×10^{-5} and batch size 16.

For all neural baselines, training is performed with early stopping on validation F1 rather than a fixed epoch budget, which typically results in fewer than 30 epochs. The best model checkpoint is used for evaluation.

D. Evaluation Metrics

Framework performance was evaluated using five standard classification metrics: Accuracy, Precision, Recall, F1-score, and the Area Under the ROC Curve (AUC). Accuracy measures the overall correctness of predictions, while Precision and Recall capture the model’s ability to identify true causal links without introducing false positives or missing actual relations. The F1-score, as their harmonic mean, provides a balanced indicator of model performance under class imbalance. AUC quantifies the trade-off between true positive and false positive rates, reflecting the model’s discriminative capability across different thresholds.

E. Results

1) *Baseline Comparison:* Table I summarizes the performance of CaST and existing baselines on the disaster tweet dataset. Overall, CaST achieves the best results across most metrics, attaining 0.87 accuracy, 0.84 precision, 0.85 recall, and 0.85 F1-score. These consistent gains confirm the effectiveness of incorporating spatial and temporal signals into event representations for causal discovery.

TABLE I
PERFORMANCE COMPARISON ON THE IN-HOUSE DATASET.

Model	Acc.	Prec.	Rec.	F1	AUC
ANM [28]	0.50	0.50	0.53	0.52	0.50
RF	0.71	0.65	0.69	0.65	0.68
SVM	0.63	0.63	0.68	0.60	0.68
BiLSTM+Att. [29]	0.70	0.60	0.62	0.61	0.65
BERT [30]	0.79	0.74	0.81	0.75	0.89
DECI [11]	0.54	0.62	0.79	0.70	0.58
PPAT [17]	0.86	0.60	0.30	0.40	0.88
DAPrompt [31]	0.77	0.52	0.78	0.62	0.86
CaST (Ours)	0.87	0.84	0.85	0.85	0.85

Among the baselines, transformer-based methods such as BERT and DAPrompt outperform earlier neural and statistical approaches. BERT achieves an F1-score of 0.75 and the highest AUC of 0.89, indicating its strong semantic understanding of contextual relations between event pairs. However, it lacks explicit modeling of event occurrence in space and time, which limits its ability to capture causal propagation patterns typical of disaster scenarios. DAPrompt performs comparably (F1 = 0.62, AUC = 0.86) due to its use of adaptive prompting, but its precision remains low, suggesting frequent false positives when causal cues are implicit or ambiguous.

Graph-based models, including DECI and PPAT, demonstrate mixed performance. DECI achieves high recall (0.79)

but low accuracy (0.54), reflecting its tendency to over-predict causal connections. PPAT, conversely, attains the highest baseline accuracy (0.86) but at the cost of poor recall (0.30), indicating that its progressive refinement captures only clear, surface-level causality while missing more subtle links. Traditional machine learning baselines (SVM, RF) perform moderately well, showing that lexical and syntactic cues can offer limited discriminative power, while ANM performs near random due to its reliance on numerical assumptions unsuitable for textual input. Importantly, all of the baselines above hardly achieve a balanced trade-off between precision and recall. Most of them tend to favor one over the other, either missing critical causal links or introducing excessive false positives. This imbalance indicates that existing methods often overlook important causal information dispersed across space and time, leading to incomplete understanding of event propagation in disasters.

In contrast, CaST effectively balances both semantic and structural reasoning. By embedding spatial and temporal features directly into event representations before graph learning, it captures dependencies that span across locations and time intervals—an aspect particularly relevant in disaster contexts where cascading effects often follow a spatial-temporal chain (e.g., flooding leading to power outages). This integration enables CaST to reduce false positives compared to BERT and improve recall compared to PPAT, achieving robust generalization on real-world disaster narratives. Quantitatively, CaST attains both high precision (0.84) and recall (0.85), demonstrating a well-balanced capability to identify true causal links without over-predicting non-causal ones. This balance is particularly crucial for disaster response, where false positives can mislead situational understanding and false negatives may overlook critical event dependencies. These results highlight that causal reasoning in social media text benefits significantly from structured spatio-temporal context rather than relying solely on semantic cues or static graph reasoning. CaST’s improvements demonstrate the value of unifying spatial, temporal, and semantic representations within a single graph-based learning framework for disaster-related causal discovery.

It is worth noting that the dataset is inherently imbalanced, with causal event pairs forming only a small fraction of all possible event combinations (approximately 1:3 in this in-house dataset). As a result, several baseline models exhibit relatively high accuracy but noticeably lower precision, recall, and F1 scores. This occurs because such models are prone to the imbalance in the data, tending to predict the majority non-causal class more frequently. Consequently, accuracy alone can be misleading in this setting, while F1 and AUC serve as more reliable indicators of causal detection performance, as they better reflect the model’s ability to identify true causal links amid class imbalance. To further examine the contribution of each component in CaST, we conduct an ablation study that isolates the effects of spatial and temporal features, as discussed in the next subsection.

F. Ablation Study

To evaluate the contribution of spatial and temporal reasoning in CaST, we perform a series of ablation experiments and compare four configurations: without both spatial and temporal components (text-only), without spatial, without temporal, and the full model. The quantitative results are summarized in Table II, while the corresponding learning trends are shown in Figure 4 and Figure 5.

TABLE II
COMPARISON OF CAST VARIANTS ON THE IN-HOUSE DATASET.

Configuration	Acc	Prec.	Recall	F1	AUC
w/o Spatial & Temporal	0.81	0.79	0.78	0.79	0.81
w/o Temporal	0.83	0.80	0.79	0.80	0.83
w/o Spatial	0.84	0.80	0.78	0.79	0.82
Full (CaST)	0.87	0.84	0.85	0.85	0.85

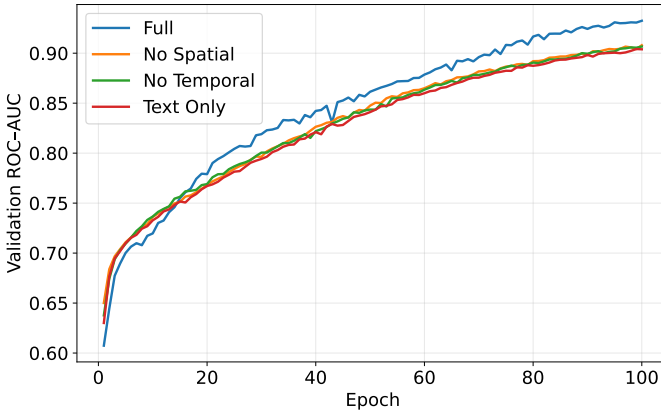


Fig. 4. ROC-AUC comparison of CaST variants during training.

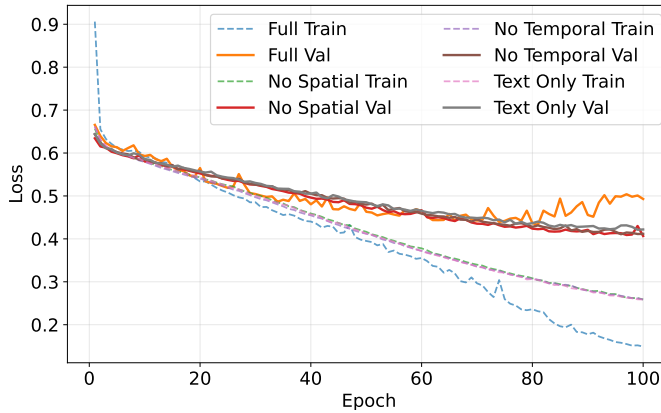


Fig. 5. Training and validation loss curves for CaST and its ablated variants.

Ablation results. The base variant without either component (*w/o Spatial & Temporal*) shows the lowest performance across all metrics ($F1 = 0.79$), underscoring that semantic embeddings alone are insufficient for representing the dynamic and context-dependent nature of disaster events. Introducing spatial or temporal information individually yields measurable

improvements of about 1–2% across most metrics, indicating that each dimension contributes complementary information to causal reasoning. The model without temporal information (*w/o Temporal*) performs slightly worse than the one without spatial cues (*w/o Spatial*), which aligns with the principle that causal relationships often follow a chronological sequence—events that occur earlier are more likely to cause later ones. Temporal modeling thus enhances recall by identifying ordered event dependencies, while spatial modeling strengthens precision by filtering unrelated events that may co-occur temporally but occur in distinct geographic regions.

When both spatial and temporal dimensions are integrated, the full CaST model achieves the highest overall performance ($Acc = 0.87$, $F1 = 0.85$, $AUC = 0.85$), showing a balanced trade-off between precision and recall. This balanced improvement again highlights the synergy between spatial proximity and temporal progression: together they enable the model to capture realistic causal propagation patterns that unfold over space and time during disasters. These findings affirm that causal reasoning in the disaster domain cannot rely solely on semantic cues or pairwise correlations but must jointly encode spatio-temporal dependencies to achieve robust and interpretable causal discovery.

Training Stability and Generalization Trends. Figure 4 compares the validation ROC-AUC of all configurations across training epochs. The full CaST model consistently outperforms its ablated counterparts, with the performance gap widening steadily after around 40 epochs. This trend indicates that spatial and temporal cues provide complementary context that strengthens causal discrimination as training progresses. In contrast, models lacking these features tend to plateau earlier, suggesting limited capacity to capture long-range or cross-event dependencies. To further understand this behavior, Figure 5 examines the convergence dynamics across configurations. The full CaST variant not only achieves the lowest training loss but also converges more smoothly, while the ablated models exhibit minor oscillations, indicating overfitting and unstable optimization. This stability implies that spatio-temporal information acts as a structural regularizer, guiding the model to learn more coherent event dependencies across time and space.

V. CONCLUSION

This paper presented CaST, a novel framework for causal discovery via spatio-temporal graphs in disaster-related tweets. Motivated by the need to understand how disaster effects propagate across regions and time, CaST unifies linguistic, spatial, and temporal cues into a single event-centric graph structure. The framework encodes contextual, temporal, and location-aware dependencies between events extracted from social media messages and learns to infer causal relations through a multi-head Graph Attention Network (GAT). Unlike prior text-only or static graph approaches, CaST explicitly models the dynamic and multi-dimensional nature of disaster events. Each event representation integrates semantic embeddings with temporal ordering and geolocation features,

allowing the network to reason over evolving causal chains. The incorporation of focal loss further enhances the model’s robustness under severe class imbalance, emphasizing hard-to-detect causal links that are often overlooked in large-scale social datasets. Extensive experiments on our in-house dataset of 167K disaster tweets show that CaST achieves consistent improvements over state-of-the-art baselines, with superior F1 and AUC scores. In future work, we aim to explore how external or domain-specific knowledge can be leveraged to enrich event representations and improve causal inference.

ACKNOWLEDGMENT

This work was supported by NSF - USA CNS-2219614.

REFERENCES

- [1] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” *International Conference on Learning Representations*, 2018, accepted as poster. [Online]. Available: <https://openreview.net/forum?id=rJXMPikCZ>
- [2] C. S. G. Khoo, J. Kornfilt, R. N. Oddy, and S. H. Myaeng, “Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing,” *Literary and Linguistic Computing*, vol. 13, no. 4, pp. 177–186, 1998.
- [3] E. Blanco, N. Castell, and D. Moldovan, “Causal relation extraction,” in *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC’08)*, 2008.
- [4] R. Girju, “Automatic detection of causal relations for question answering,” in *Proc. of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, 2003, pp. 76–83.
- [5] P. Cao, X. Zuo, Y. Chen, K. Liu, J. Zhao, Y. Chen, and W. Peng, “Knowledge-enriched event causality identification via latent structure induction networks,” in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, 2021, pp. 4476–4488.
- [6] S. Shen, H. Zhou, T. Wu, and G. Qi, “Event causality identification via derivative prompt joint learning,” in *Proc. of the 29th International Conference on Computational Linguistics (COLING 2022)*, 2022, pp. 2288–2299.
- [7] X. Zuo, P. Cao, Y. Chen, K. Liu, J. Zhao, W. Peng, and Y. Chen, “LearnDA: Learnable knowledge-guided data augmentation for event causality identification,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 3558–3571.
- [8] S. Anuyah, J. Vanschaik, P. Jain, S. Lehman, and S. Chakraborty, “An empirical study of causal relation extraction transfer: Design and data,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.06076>
- [9] Y. Chun, S. Ha, J. Bai *et al.*, “Utilizing pre-trained language models for data augmentation task of event causality identification,” *Research Square*, May 2025, preprint (Version 1).
- [10] L. Gao, P. K. Choubey, and R. Huang, “Modeling document-level causal structures for event causal relation identification,” in *Proc. of NAACL-HLT 2019 (Volume 1: Long Papers)*, 2019, pp. 1808–1817.
- [11] M. Tran Phu and T. H. Nguyen, “Graph convolutional networks for event causality identification with rich document-level structures,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 3480–3490.
- [12] K. Fan, P. Li, Y. Du, D. Yu, Z. Yu, Y. Zhang, and W. Lu, “Document-level event causality identification via heterogeneous graph attention networks,” in *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, 2022, pp. 1252–1263.
- [13] C. Liu, W. Xiang, and B. Wang, “Identifying while learning for document event causality identification,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3815–3827. [Online]. Available: <https://aclanthology.org/2024.acl-long.210/>
- [14] R. Pu, Y. Li, J. Zhao, S. Wang *et al.*, “A joint framework with heterogeneous-relation-aware graph and multi-channel label enhancing strategy for event causality extraction,” in *Proc. of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024)*, 2024, pp. 18 879–18 887.
- [15] Z. Hu, Z. Li, X. Jin, L. Bai, S. Guan, J. Guo, and X. Cheng, “Semantic structure enhanced event causality identification,” in *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 2023, pp. 10 901–10 913.
- [16] P. Mirza and S. Tonelli, “An analysis of causality between events and its relation to temporal information,” in *Proc. of COLING 2014 (25th International Conference on Computational Linguistics)*, 2014, pp. 2097–2106.
- [17] Z. Liu, B. Hu, Z. Xu, and M. Zhang, “Ppat: progressive graph pairwise attention network for event causality identification,” in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, ser. IJCAI ’23, 2023.
- [18] Z. Liu, Y. Liang, and W. Ni, “Document-level causal event extraction enhanced by temporal relations using dual-channel neural network,” *Electronics*, vol. 14, no. 5, p. 992, 2025.
- [19] X. Dong, E. Mas, B. Adriano, and S. Koshimura, “Towards real-time extraction of cascading effect and spatiotemporal analysis using social media data,” *International Journal of Disaster Risk Reduction*, vol. 125, p. 105512, 2025.
- [20] J. Lenti *et al.*, “Causal modeling of climate activism on reddit,” in *Proc. of The Web Conference (WWW) 2025*, 2025, pp. 590–600.
- [21] E. Kiciman, R. Ness, A. Sharma, and C. Tan, “Causal reasoning and large language models: Opening a new frontier for causality,” *Transactions on Machine Learning Research (TMLR)*, 2024, to appear; arXiv:2305.00050.
- [22] S. Long, A. Piché, V. Zantedeschi, T. Schuster, and A. Drouin, “Causal discovery with language models as imperfect experts,” 2023.
- [23] C. Shyalika and *, “Causal event graph-guided language-based spatiotemporal question answering,” in *Proc. of the AAAI Spring Symposium on Knowledge Graphs for XAI*, 2024.
- [24] P. W. Holland, “Statistics and causal inference,” *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.
- [25] R. Lamsal, M. R. Read, and S. Karunasekera, “Crisistransformers: Pre-trained language models and sentence encoders for crisis-related social media texts,” *Knowledge-Based Systems*, vol. 296, p. 111916, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705124005501>
- [26] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *CoRR*, vol. abs/1708.02002, 2017. [Online]. Available: <http://arxiv.org/abs/1708.02002>
- [27] X. Wang, Y. Chen, N. Ding, H. Peng, Z. Wang, Y. Lin, X. Han, L. Hou, J. Li, Z. Liu, P. Li, and J. Zhou, “Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction,” in *Proceedings of EMNLP*, 2022.
- [28] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, “Nonlinear causal discovery with additive noise models,” in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2008.
- [29] Z. Li, Q. Li, X. Zou, and J. Ren, “Causality extraction based on self-attentive bilstm-crf with transferred embeddings,” *CoRR*, vol. abs/1904.07629, 2019.
- [30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [31] W. Xiang, C. Zhan, Q. Zhang *et al.*, “Daprompt: deterministic assumption prompt learning for event causality identification,” *Neural Computing and Applications*, vol. 37, pp. 21 743–21 759, 2025.