EPIR: Capturing Promoting and Inhibiting Relationships between Events

Bowen Dong[1], Wenjun Wang[1], Xueli Liu[2]*, Quanlin Qiu[3],
[1]School of Artificial Intelligence, Tianjin University
[2]School of CyberSecurity, Tianjin University
[3]School of Computer Science and Technology, Tianjin University
{1020201107,wjwang, xueli, tolalal}@tju.edu.cn

Abstract

Understanding whether one event increases or decreases the likelihood of another is critical for real-life applications. Unlike other relationships, promoting and inhibiting relationships capture directional, probabilistic, and context-dependent shifts in event likelihood. A central challenge is to estimate this relative influence from observational data: naive conditional probabilities conflate influence with correlation and are easily distorted by shared contextual confounders. We propose EPIR, a unified framework for estimating promoting and inhibiting relationships from observed event data. EPIR formulates influence as a relative directional effect under comparable contextual conditions, and models event context using : (i) observable history captured and (ii) latent multi-hop propagation mechanisms. EPIR combines context-conditioned predictive evidence with schema-based structural evidence to produce a single signed influence score, where the sign determines promotion versus inhibition. Experiments on real-world datasets show that EPIR outperforms other baselines in accuracy. EPIR is available at https://github.com/EveLappland/EPIR.git.

1 Introduction

Understanding how events influence each other is fundamental to analyzing the dynamics of complex systems such as international conflicts, financial crises, public health emergencies, and online discourse (Wang et al., 2020; Do et al., 2012). Beyond identifying what events occur and when they occur, many real-world decision-making tasks depend on understanding whether one event *increases or decreases the likelihood* of another.

We focus on such directional influences, which are critical in domains where risk accumulates over time and early signals matter. In public safety and disaster management, identifying events that promote escalation or inhibit recovery supports early warning and intervention planning. In economic and social systems, understanding which signals amplify instability or dampen collective responses enables more robust policy design and risk mitigation. More broadly, promoting and inhibiting relationships provide a principled way to model how complex systems evolve under uncertainty, making them essential for prediction, monitoring, and strategic decision-making.

A key challenge is that promoting and inhibiting influence is not an absolute property of an event pair, but a relative effect: it depends on how the likelihood of a downstream event changes when a preceding event occurs versus when it does not. This effect is context-dependent, shaped by historical, environmental, and semantic conditions, while typically propagating through multi-step event chains rather than a single direct link.

Example 1: An illustrative example of *promoting influence* is the 2021 Texas winter storm, where extreme cold, under state-specific infrastructure and regulatory conditions, triggered cascading power outages by amplifying latent vulnerabilities along a multi-step failure chain rather than acting as a deterministic cause.

An example of *inhibiting influence* arises in wildfire prevention. In Michigan, wind-based burn alerts reduced human-caused wildfire ignitions during high-wind periods compared with similar conditions before the policy, showing how persistent adverse environments can be mitigated by interrupting intermediate behavioral mechanisms. □

These characteristics distinguish promoting and inhibiting relationships from traditional event associations such as co-occurrence (Benezeth et al., 2009), temporal dependencies (Wu et al., 2024; Han et al., 2019), and causal chains (Cao et al., 2021; Bhattacharjya et al., 2021; Han et al., 2019;

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36895–36910

Cai et al., 2013; Li and Liu, 2022). Co-occurrence captures undirected correlation without influence; temporal order reflects sequence but not impact; and causal analysis seeks necessary or sufficient conditions under strong assumptions that rarely hold in open event streams. In contrast, promoting and inhibiting relationships capture probabilistic, context-dependent shifts in likelihood, without requiring deterministic causality.

A natural approach to modeling influence is to rely on conditional probabilities. However, simple measures such as $P(v_j \mid v_i)$ conflate directional influence with correlation and provide no counterfactual baseline. A classic example is the observed correlation between ice cream sales and drowning incidents: although $P(\text{drowning}|\text{ice cream})$ is high, ice cream does not promote drowning, both are driven by a latent confounder, namely hot weather.

Accurately modeling promoting and inhibiting relationships therefore requires: (i) defining influence as a *relative change* in probability rather than an absolute conditional likelihood; (ii) grounding this comparison in a *sufficient event context* to control for shared factors; and (iii) accounting for *latent, multi-hop mechanisms* through which influence propagates across events.

Contributions. This paper makes the following:

(1) Directional Influence Modeling. We define promoting and inhibiting relationships as context-dependent directional influences, measured by the relative change in the likelihood of a downstream event conditioned on comparable contexts.

(2) Unified Relationship Analysis Framework. We propose EPIR, a unified framework that estimates such directional influences from observational event data by jointly modeling observable semantic signals and latent dependency patterns.

(3) Experimental Results. We conduct experiments on real-world event datasets and find the following: (a) EPIR is effective, outperforming other approaches by 2%, 10%, and 5% on these metrics, respectively. (b) Ablation studies indicate consistent performance gains from each EPIR component.

2 Related Work

Event Association Analysis. Prior studies can be broadly categorized into temporal relations, subevent relations, and causal relations. Temporal relation extraction focuses on identifying ordering relations based on their temporal attributes (Han et al., 2019; Cai et al., 2013). Recent work has enhanced temporal modeling by incorporating logic-based representations of event boundaries and durations (Huang et al., 2023). Subevent relation analysis aims to uncover hierarchical structures between events (Wang et al., 2020). Causal event analysis seeks to identify directional dependencies between events, often relying on assumptions such as pairwise co-occurrence (Luo et al., 2016; Granger, 1969), statistical correlation (Peters et al., 2017), conditional intensity modeling (Bhattacharjya et al., 2021), or probabilistic graphical models (Bhattacharjya et al., 2020). More recent approaches explore centrality-aware causal structures (Hu et al., 2023), counterfactual reasoning (Mu and Li, 2023), and event interaction graphs (Tao et al., 2023).

3 Contextual Influence Modeling via Evidence Decomposition

3.1 World Context beyond Event History

Existing studies on event influence often equate context with event history, implicitly assuming that past events fully characterize the conditions under which future events occur. This assumption is insufficient in open-world settings, where event occurrences are shaped not only by external environmental factors (e.g., economic conditions), but also by structural constraints (e.g., causal mechanisms) that determine which transitions are feasible

To capture these heterogeneous factors in a unified manner, we model context $\mathbf{C}$ as a structured composition of three complementary components: $(C_{\text{hist}},\ C_{\text{env}},\ C_{\text{struct}})$ where: (1) C_{hist} represents the event history, including past events and their temporal organization; (2) C_{env} represents exogenous environmental conditions, such as economic or political factors; and (3) C_{struct} represents structural constraints, including institutional rules or domain-specific mechanisms that govern feasible event transitions.

3.2 Influence Model

We first propose formal definitions of events.

Events. An event is represented as a tuple $v = (t, l, L, S)$, where t denotes the time of occurrence, l denotes the location associated with the event, L is a discrete event type label, and $S = (s, p, o)$ is a triple describing the interaction between a source entity s, a predicate p, and a target entity o.

Given two events v_i and v_j, our goal is to de-

termine whether v_i promotes or inhibits the occurrence of v_j under a fixed context $\mathbf{C}$. Formally, the ideal influence quantity is the signed contrast:

$$\Delta P = P(v_j \mid v_i, \mathbf{C}) - P(v_j \mid \neg v_i, \mathbf{C}), \quad (1)$$

which captures the directional effect of v_i on v_j.

Unified Treatment of Promotion and Inhibition. The definitions of promoting and inhibiting relationships are directly grounded in the contrast ΔP. Under a fixed context, ΔP measures the change in the tendency of an event to occur when another event is present versus absent. A positive value of ΔP indicates that the presence of the event increases this tendency, corresponding to promotion, while a negative value indicates a decrease, corresponding to inhibition. Thus, promotion and inhibition naturally arise as opposite signs of the same contrast quantity and can be computed in a unified manner without introducing separate models.

From a computational perspective, directly estimating ΔP is challenging due to the partial observability of the world context. We therefore infer both the direction and the strength of the influence using context-aware evidence functions constructed over matched world contexts.

3.3 Evidence Decomposition

A key design choice in our framework is that not every component of $\mathbf{C}$ induces a distinct evidence type. Instead, the three context components naturally combine into two complementary forms of evidence with different inferential roles.

Specifically, (1) whether an event *will occur* is primarily determined by past event trajectories (C_{hist}) together with external environmental conditions (C_{env}). (2) whether an event *should occur because of another event* is governed by environmental triggers (C_{env}) and structural mechanisms or constraints (C_{struct}).

Accordingly, we construct two evidence types: (1) Observational Predictive Evidence $(C_{\text{hist}}, C_{\text{env}})$; (2) Mechanistic Coherence Evidence $(C_{\text{env}}, C_{\text{struct}})$. The environmental context C_{env} serves as a bridging variable, influencing both empirical predictability and mechanistic validity.

Observational Predictive Evidence. We first define evidence that captures whether the presence of v_i empirically increases the tendency for v_j to occur under comparable world conditions.

Observed World State. We define an observed world state at time t as $O_t = (E_{\leq t}, X_t)$, where $E_{\leq t}$ corresponds to the event history (C_{hist}) and X_t corresponds to environmental variables (C_{env}).

Predictive Scoring Function. Let $\phi_{\text{obs}}(O_t, v_j)$ denote a learned scoring function that measures the predictive consistency between the observed world state O_t and the occurrence of event v_j. This score is not required to be a calibrated probability; it only needs to be monotonically correlated with the likelihood of v_j.

Context-Matched Contrast. To isolate the effect of v_i, we construct two matched sets of observed states: (1) $\mathcal{O}^+$: states containing v_i; (2) $\mathcal{O}^-$: states not containing v_i, where states in both sets are matched with respect to environmental conditions.

The observational predictive evidence $E_{\text{obs}}(v_i \rightarrow v_j)$ (shortened as E_{obs}) is defined as:

$$\mathcal{E}_{O \in \mathcal{O}^+}[\phi_{\text{obs}}(O, v_j)] - \mathcal{E}_{O \in \mathcal{O}^-}[\phi_{\text{obs}}(O, v_j)]\,.$$

A positive value indicates that, under comparable world conditions, v_i is associated with a higher estimated likelihood for v_j to occur.

Mechanistic Coherence Evidence. Observational patterns alone cannot distinguish genuine influence from spurious correlations. We therefore introduce a second form of evidence that evaluates whether an event relation is mechanistically plausible.

Mechanistic Context. We define the mechanistic context as $M = (C_{\text{env}}, C_{\text{struct}})$, capturing both environmental triggers and structural constraints.

Mechanism Scoring. Let $\phi_{\text{mech}}(v_i \rightarrow v_j \mid M)$ denote a scoring function that evaluates how strongly the event pair (v_i, v_j) is supported by latent schemas or mechanisms under context M.

Evidence Aggregation. The mechanistic coherence evidence $E_{\text{mech}}(v_i \rightarrow v_j)$ is obtained by softly aggregating schema-level mechanism scores:

$$E_{\text{mech}} = \text{log-sum-exp}\Big(\phi_{\text{mech}}(v_i \rightarrow v_j \mid M)\Big),$$

where each $\phi_{\text{mech}}(\cdot)$ reflects the confidence of an individual latent dependency schema.

Contextual Influence Estimation. Finally, we combine the two evidence types to approximate the signed influence contrast:

$$\Delta_{\text{ctx}}(v_i, v_j) = f\Big(E_{\text{obs}}(v_i \rightarrow v_j),\ E_{\text{mech}}(v_i \rightarrow v_j)\Big),$$

where $f(\cdot)$ is a fusion function that balances empirical predictability and mechanistic coherence.

We determine the type of directional influence by thresholding Δ_{ctx}: values above τ indicate promotion, values below $-\tau$ indicate inhibition, and

intermediate values are treated as insignificant. We next instantiate the predictive and mechanistic scoring functions using neural representation learning and schema-based reasoning, respectively.

Discussion. We discuss how EPIR distinguishes promoting relations from subevent relations. In EPIR, promoting relations denote probabilistic influence (i.e., whether one event changes the occurrence tendency of another under matched environments), whereas subevents represent structural containment. The dual-evidence design avoids conflation: observational evidence requires a stable ΔP under matched contexts, and structural evidence requires cross-chain reusable mechanisms. Pure compositional relations typically lack such ΔP and transferable support, and are therefore not identified as promoting relations.

4 EPIR:A Predictive Framework for Influence Inference

This section presents a predictive framework for computing contextual influence between events.

Framework Overview. The framework operates by constructing a unified representation of observable world context and deriving influence scores from two complementary sources of evidence. First, observational evidence is computed from event sequences by modeling how the presence of one event alters the predictive tendency of another under matched contextual conditions. Second, structural evidence is obtained by mining recurrent transition mechanisms from historical data and evaluating their consistency under the current environmental context. These two evidence signals capture different aspects of influence: empirical predictability and mechanistic stability.

4.1 Observed Evidence computation

Modeling Temporal History via Event Chains. We model historical context as a structured temporal object that captures how past events jointly constrain future event tendencies. Rather than treating history as a fixed-length window or an unordered set of events, we represent temporal history as an event chain whose internal structure is learned to encode predictive dependencies.

Formally, given a sequence of timestamped events $E_{\leq t} = \{v_1, v_2, \ldots, v_k\}$, we define a temporally ordered chain $\mathcal{C}_t = (v_1 \prec v_2 \prec \cdots \prec v_k)$, where $\prec$ denotes the temporal order induced by event timestamps, i.e., $v_i \prec v_j$ if $v_i.t_i < v_j.t_j$. The detailed construction steps are illustrated in Appendix A.2

The event chain representation encodes event order, relative temporal structure, and semantic composition across events, forming a latent temporal state that constrains future tendencies. Specifically, to encode $\mathcal{C}_t$, we employ a Mixed-Context Reasoning Module (Detailed description are shown in Appendix A.2) that maps the entire event chain into a latent temporal state $h_{\mathcal{C}} = \textsc{ChainEncode}(\mathcal{C}_t)$. The encoder adopts channel-wise and patch-wise mixing to achieve a global receptive field, allowing temporally distant yet semantically related events to directly interact. This latent state summarizes predictive constraints imposed by the full event history, rather than by individual events in isolation.

Environmental Context. Environmental context is modeled as an external process that evolves over time independently of the event chain. Rather than assuming a static environment, we explicitly represent environmental information as a time-indexed signal aligned with the event timeline.

Time-Indexed Environmental Representation. We represent the environmental context as time-indexed observations $\mathcal{X} = \{X(t) \mid t \in \mathcal{T}\}$, where each $X(t) \in \mathcal{R}^d$ denotes a vector of environmental attributes observed at time t. The temporal resolution of $\mathcal{X}$ may differ from that of the event sequence. In EPIR, environmental signals are aligned to event timestamps and used to condition event representations, detailed event representations are shown in appendix A.2.

Environmental State Encoding. We encode the aligned environmental observations up to time t into a compact representation: $\mathbf{h}_X = \phi_{\text{env}}(\mathcal{X}_{\leq t})$, where $\phi_{\text{env}}(\cdot)$ denotes a lightweight environmental encoder that aggregates time-indexed environmental attributes aligned to the event timestamps. In EPIR, we encode environmental signals by pooling observations aligned to event timestamps. The aggregated representation is then concatenated with the event-chain representation as the final context.

Full Context Representation. We construct the full observable context representation by integrating the event-chain representation and the aligned environmental encoding $h_O = h_{\mathcal{C}} \oplus h_X$, where $\oplus$ denotes vector concatenation. The resulting representation h_O serves as unified description of the observable world context at time t.

Directional Influence Scoring. Given two events

v_i and v_j with embeddings h_{v_i} and h_{v_j}, we compute a signed influence score by comparing the predicted tendency of v_j under two contextualized states: one that includes v_i as an active factor and one that excludes it. Specifically, We define

$$z^{+}_{i\to j} = s\big(h_O \oplus h_{v_i},\, h_{v_j}\big)\,, \tag{2}$$

$$z^{-}_{i\to j} = s\big(h_O \oplus \mathbf{0},\, h_{v_j}\big)\,, \tag{3}$$

$$S_{\text{obs}}(v_i \to v_j) = \sigma\Big(z^{+}_{i\to j}\Big) - \sigma\Big(z^{-}_{i\to j}\Big)\,, \tag{4}$$

where $s(\cdot,\cdot)$ is implemented by cosine similarity and $\sigma(\cdot)$ is the squashing function. Here $h_O \oplus h_{v_i}$ represents the historical context explicitly conditioned on event v_i, while $h_O \oplus \mathbf{0}$ serves as an unconditioned baseline. The resulting $S_{\text{obs}}(v_i \to v_j)$ is a signed predictive signal: positive values indicate promotion, and negative values indicate inhibition.

4.2 Structural Evidence Computation

This subsection describes how such structural mechanisms are mined from data and how their consistency is quantified under the current environmental conditions.

Structural Mechanism Space. We represent structural context as a set of abstract mechanisms, each encoding a reusable transition pattern between event types. Formally, a structural mechanism m is defined as a triple (v_i, v_j, ψ) where (v_i, v_j) specifies a directed event transition, and ψ denotes a set of environmental applicability conditions under which the transition is considered valid.

Mechanism Mining from Event Data. Structural mechanisms are mined from historical event data by inducing recurrent transition structural mechanisms s_k at the event-type level, which capture latent structural dependencies beyond direct event co-occurrences. Rather than relying on frequent instance-level transitions, we abstract event sequences into type-level patterns whose conditional support persists across multiple event chains within a temporal window. These mechanisms characterize multi-hop mechanisms that may not be explicitly observable in any single event chain.

During inference, candidate mechanisms are evaluated under the current historical context by matching their structures against event paths anchored at v_i. The output of this process is a collection of latent mechanisms $\mathcal{M} = \{s_k\}$, where each mechanism s_k encodes a multi-hop transition pattern at the event-type level and is associated with a learned confidence score that reflects its global reliability and contextual applicability.

Discussion. In EPIR, the mechanism set $\mathcal{M}$ is mined offline and remains fixed during training and inference; no mechanisms are created at runtime. Only type-level patterns meeting predefined support, confidence, and stability thresholds are retained, while others are discarded. LLM-generated candidates are treated as hypotheses and must pass the same statistical validation before inclusion. The process is fully automated and does not rely on human inspection.

LLM-Augmented Mechanism Generation. Purely data-driven mechanism mining is often sparse in open-world event datasets, as many meaningful structural dependencies occur infrequently or are only partially observed. To mitigate this sparsity, we augment core mechanisms with an LLM-assisted generation step, while explicitly controlling for hallucinated structures.

For mechanism augmentation, the LLM is prompted with a structured input consisting of (i) the set of event types in the domain, (ii) environmental conditions represented as categorical attributes, and (iii) a small number of example mechanisms illustrated by concrete instances. Conditioned on this input, the LLM generates candidate environment-conditioned mechanism hypotheses describing how event types may interact under specific environments.

Each generated hypothesis specifies a type-level mechanism s_k, accompanied by a qualitative rationale and illustrative examples. The rationale and examples are used exclusively for human inspection to discard hallucinated or implausible mechanisms, while only the verified type-level mechanism structures are retained as candidate mechanisms $\mathcal{M}_{\text{LLM}}$ for subsequent inference.

Environment-Conditioned Mechanism Activation. Given a candidate pair (v_i, v_j) and the current historical context, EPIR activates a mechanism $s_k \in \mathcal{M}$ by matching it to event paths anchored at v_i. Here, $\mathcal{M}$ includes both data-mined core mechanisms and LLM-augmented mechanisms that have been verified through human inspection. For each mechanism, we compute a context alignment score by aggregating the matching quality over all paths consistent with s_k:

$$a_k = c_k \cdot \sum_{p \in \mathcal{P}(v_i, s_k)} c_{\text{ctx}}(p), \tag{5}$$

where c_k denotes the global confidence of s_k and $c_{\text{ctx}}(p)$ measures how well a matched path p aligns with the current environment and the mechanism's applicability conditions. mechanisms with low confidence are naturally downweighted.

Structural Consistency Scoring. For a candidate transition (v_i, v_j), we compute a structural consistency score by aggregating the contributions of all activated mechanisms that support this transition:

$$S_{\text{struct}}(v_i \to v_j) = \sum_{m_k \in \mathcal{M}_{i,j}} a_k \cdot w_k, \qquad (6)$$

where $\mathcal{M}_{i,j}$ denotes the subset of mechanisms associated with (v_i, v_j), and w_k is a learned or data-driven weight reflecting the reliability of mechanism m_k. This score quantifies the degree to which the transition is structurally supported by consistent mechanisms under the current environment.

4.3 Evidence Fusion Computation

Given the observational predictive score and the structural consistency score computed in the previous subsections, we construct a joint evidence model to infer the overall influence between two events. This subsection specifies how heterogeneous evidence signals are integrated into a unified, signed influence score.

Evidence Normalization. Since the observational and structural evidence scores are produced by different modules and reside on different scales, we normalize them to ensure comparability before fusion. For an event pair (v_i, v_j), we obtain normalized evidence signals:

$$\tilde{S}_{\text{obs}}(v_i \to v_j) = \sigma(S_{\text{obs}}(v_i \to v_j)), \qquad (7)$$

$$\tilde{S}_{\text{struct}}(v_i \to v_j) = \sigma(S_{\text{struct}}(v_i \to v_j)), \qquad (8)$$

where $\sigma(\cdot)$ denotes the sigmoid function. This normalization maps heterogeneous evidence scores into a unified $[0, 1]$ range.

Fusion Model. We integrate the two normalized evidence signals using a fusion function:

$$\Delta_{ctx}(v_i, v_j) = \textsc{Fuse}\big(\tilde{S}_{\text{obs}}, \tilde{S}_{\text{struct}}\big), \qquad (9)$$

where $\textsc{Fuse}(\cdot)$ is implemented as a multilayer perceptron (MLP).

According to the definitions in Section 3, the sign and magnitude of $\Delta_{ctx}(v_i, v_j)$ are then used to predict promoting or inhibiting relationships.

4.4 Trainable Models

Trainable vs. Non-trainable Components. Let θ_{enc} denote the parameters of the event-chain encoder, environment encoder, and contextualization operator, θ_{score} denote the parameters of the state–event compatibility scorer, and θ_{fuse} denote the parameters of the fusion model. We train the parameter set $\Theta = \{\theta_{\text{enc}}, \theta_{\text{score}}, \theta_{\text{fuse}}\}$, while the structural mechanism set $\mathcal{M}$ mined by the non-parametric procedure (Section 4.2) is fixed during learning.

Training Signal from Observational Sequences. For each timestamp t, we form a contextual state representation $h_O(t)$ from the observed event chain and environmental trajectory. Let v^+ be an event observed to occur in a prediction horizon after t, $\mathcal{N}_t$ be a set of negative events sampled from non-occurring events under comparable contexts (the generation of negative samples are illustrated in Appendix A.2), $\phi^+_{\text{obs}} = \phi_{\text{obs}}(h_O(t), v^+)$ and $\phi^-_{\text{obs}} = \phi_{\text{obs}}(h_O(t), v^-)$. We train the compatibility scorer using a contrastive loss: $\mathcal{L}_{\text{rank}}(\Theta)$ as

$$\sum_t \left[-\log \frac{\exp\big(\phi^+_{\text{obs}}\big)}{\exp\big(\phi^+_{\text{obs}}\big) + \sum_{v^- \in \mathcal{N}_t} \exp\big(\phi^-_{\text{obs}}\big)} \right].$$

where $\phi_{\text{obs}}(h_O(t), v)$ is the model score induced by θ_{enc} and θ_{score}. This objective propagates gradients through the encoders and the scorer, aligning the latent context representation with predictive utility.

Joint Training with Structural Evidence Fusion. Given the observational score S_{obs} and the structural consistency score S_{struct} computed from the fixed mechanism set $\mathcal{M}$, the fusion model produces the final influence score $\Delta(v_i, v_j)$ (Eq. (9)). When relation labels (promote/inhibit/neutral) are available, we train θ_{fuse} (and optionally fine-tune θ_{enc} and θ_{score}) using a supervised loss:

$$\mathcal{L}_{\text{fuse}}(\Theta) = \sum_{(v_i, v_j)} \ell\Big(y_{ij},\, \Delta(v_i, v_j)\Big), \qquad (10)$$

where y_{ij} is the ground-truth relation label and $\ell(\cdot, \cdot)$ is a classification loss (e.g., cross-entropy with margins). In the absence of relation labels, θ_{fuse} can be trained by treating frequently observed transitions as positives and unobserved transitions as negatives under matched contexts, using the same form as $\mathcal{L}_{\text{rank}}(\Theta)$.

5 Experiments

We evaluate EPIR on real-world event datasets to assess its effectiveness in predicting directional influence between events. Our experiments examine performance, the contributions of observational and structural evidence, robustness to parameter variations, and the interpretability of inferred influences.

Experimental Settings. We start with the settings. *Event Datasets*. Our experiments are conducted on five real-world event datasets, including three yearly subsets of ICEWS (Boschee et al., 2015): ICEWS-15 (2015), ICEWS-16 (2016), and ICEWS-18 (2018), as well as GDELT (Leetaru and Schrodt, 2013) and OPEN-TLS (Wu et al., 2025). Their statistics are summarized in Table 1. Each dataset is split into training, validation, and test sets according to the temporal order of events, with a ratio of 7:2:1. Within each split, a preprocessing pipeline is applied to construct chains from raw event streams.

Ground-truth labels for promoting and inhibiting relationships are obtained through a human-centered annotation protocol, where large language models assist annotators by providing candidate rationales. Event pairs for which a confident human judgment cannot be reached are discarded. Detailed descriptions, event-chain preprocessing, and label annotation are provided in Appendix A.3.

Table 1: Dataset Description.

Dataset	Events	Time Span	Attributes	Description
ICEWS-15	955,350	2015	20	political events
ICEWS-16	207,613	2016	20	political events
ICEWS-18	51,921	2018	20	political events
GDELT	100,000	1979–2014	18	social events
OPEN-TLS	353	N/A	N/A	textual reports

Environmental Context. It is obtained by first aggregating events into fixed spatiotemporal units based on their timestamps and locations. For each unit, we collect corresponding environmental signals by querying publicly available sources (e.g., weather records and geopolitical reports) and map these signals to predefined categorical environmental attributes. Then C_{env} is aligned with chains by associating events with the environmental attributes of their corresponding spatiotemporal unit.

Baselines. We compare EPIR with representative baselines spanning different paradigms of event relationship modeling to ensure a comprehensive evaluation. The baselines include: (1) Graph-based methods SLF (Xu et al., 2019), SGCN (Derr et al., 2018), SIGAT (Huang et al., 2019), SDGNN (Huang et al., 2021), and SDEGNN (Chen et al., 2024); (2) Rule-based reasoning method RNNLogic (Qu et al., 2021); (3) Causal analysis approaches Cascade (Cüppers et al., 2024) and LLM-cd (Du et al., 2025); (4) LLM-based approaches GPT-4, GPT-4o (Achiam et al., 2023), DeepSeek-v3 (Liu et al., 2024), and DeepSeek-R1 (Guo et al., 2025); and (5) Local large language model DeepSeek-R1-14B (denoted as DeepSeek-R1-14B(local)) and its fine-tuned variant(denoted as DeepSeek-R1-14B(Fined-Tuned)). Detailed descriptions are provided in Appendix A.3.

Although these baselines are not originally designed to model promoting or inhibiting relationships, they all capture directional or relational dependencies between events and can be adapted to this setting under a unified evaluation protocol.

Evaluation Metrics. We evaluate performance using (**Pre**), recall (**Rec**), and F1-score (**F1**).

Parameters Setting. For the parameters in EPIR, we set the learning rate to 0.05, the batch size to 64, and the temperature ratio to 0.11. For the mixed-content reasoning module used in EPIR, we set the depth to 2. Other relevant parameters are presented in Appendix A.

Experiment Implementation. All experiments were conducted on a server equipped with an NVIDIA GeForce RTX 4090 GPU, an AMD Ryzen Threadripper PRO 5965WX CPU (24 cores, 48 threads), and 512 GB RAM. EPIR was implemented in Python. Each experiment was repeated five times, and the average results are reported.

Experiments Results. We next report the results.

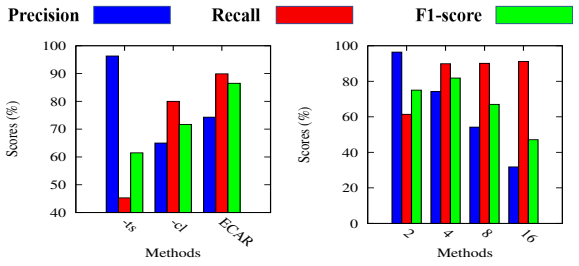
Exp-1: Accuracy. As shown in Table 2, the prediction of EPIR on all datasets achieves an average precision of 82.7%, a recall of 82.5%, and an F1-score of 82.6%. Compared with the strongest competing method on each dataset, EPIR achieves precision gains of 0.4%, 0.7%, 0.8%, 8.6%, and 4.6%, , recall gains of 6.9%, 11.6%, 8.2%, 0.3%, and 6.0% and the F1 score by 4.3%, 7.6%, 5.3%, 4.8%, and 10.1% on ICEWS-15, ICEWS-16, ICEWS-18, GDELT, and OPEN-TLS, respectively. Moreover, we add a local PLM baseline (DeepSeek-R1-14B) and its fine-tuned version. The fine-tuned model is trained on the same training split using event-chain supervision, without access to validation or test data. Results are reported in Table 2. Although fine-tuned PLMs perform strongly, EPIR consistently outperforms them across all datasets, show-

Table 2: Performance Comparison Across All Datasets. Scores are higher the better.

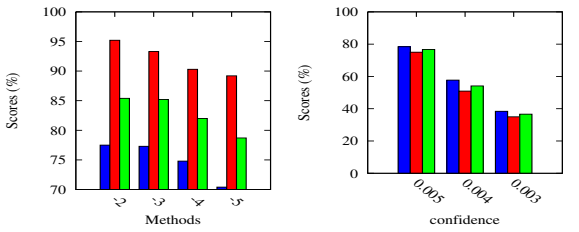
Model	ICEWS-15			ICEWS-16			ICEWS-18			GDELT			OPEN-TLS		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
SGCN	0.371	0.605	0.461	0.500	0.651	0.567	0.467	0.643	0.543	0.455	0.475	0.465	0.420	0.501	0.456
SIGAT	0.371	0.360	0.366	0.334	0.471	0.391	0.291	0.265	0.277	0.390	0.510	0.441	0.402	0.370	0.385
SDGNN	0.307	0.607	0.408	0.447	0.655	0.531	0.433	0.646	0.518	0.460	0.532	0.493	0.427	0.478	0.451
SDEGNN	0.470	0.527	0.497	0.480	0.590	0.529	0.482	0.427	0.453	0.500	0.468	0.483	0.466	0.420	0.442
SLF	0.446	0.500	0.472	0.448	0.573	0.464	0.397	0.276	0.326	0.453	0.578	0.508	0.426	0.479	0.451
RNNLogic	0.343	0.479	0.399	0.369	0.500	0.425	0.425	0.479	0.451	0.447	0.442	0.444	0.500	0.625	0.556
Cascade	0.514	0.317	0.392	0.521	0.263	0.350	0.465	0.353	0.401	0.571	0.444	0.500	0.623	0.655	0.639
LLM-cd	0.817	0.702	0.755	0.817	0.659	0.730	0.778	0.665	0.717	0.783	0.608	0.684	0.683	0.613	0.646
GPT-4	0.423	0.201	0.272	0.439	0.101	0.164	0.429	0.221	0.292	0.500	0.333	0.400	0.564	0.507	0.534
GPT-4o	0.514	0.317	0.492	0.521	0.263	0.350	0.465	0.353	0.402	0.571	0.444	0.499	0.623	0.655	0.638
Deepseek-v3	0.783	0.573	0.658	0.791	0.530	0.634	0.747	0.503	0.601	0.762	0.556	0.643	0.645	0.598	0.620
DeepSeek-R1-14B (Local)	0.423	0.520	0.467	0.431	0.483	0.456	0.458	0.480	0.469	0.547	0.553	0.550	0.533	0.504	0.518
DeepSeek-R1-14B (Fine-Tuned)	0.600	0.640	0.619	0.690	0.710	0.700	0.638	0.670	0.654	0.647	0.662	0.654	0.633	0.704	0.667
Deepseek-R1	0.850	0.830	0.840	0.842	0.787	0.813	0.809	0.827	0.818	0.803	0.659	0.724	0.720	0.627	0.670
EPIR(-LLM)(-Obs)	0.774	0.854	0.812	0.760	0.858	0.806	0.724	0.864	0.789	0.780	0.609	0.684	0.391	0.563	0.461
EPIR(-LLM)(-Struct)	0.747	0.697	0.721	0.733	0.701	0.715	0.697	0.707	0.698	0.694	0.614	0.652	0.536	0.583	0.558
EPIR(-LLM)(-temporal)	0.740	0.833	0.784	0.774	0.827	0.800	0.787	0.824	0.804	0.742	0.541	0.626	0.428	0.563	0.486
EPIR(-LLM)(-spatial)	0.700	0.850	0.770	0.732	0.844	0.783	0.745	0.841	0.790	0.464	0.270	0.342	—	—	—
EPIR(-LLM)(-semantic)	0.698	0.843	0.764	0.730	0.838	0.780	0.742	0.833	0.785	0.846	0.458	0.594	0.409	0.563	0.474
EPIR(-LLM)	0.834	0.899	0.865	0.820	0.903	0.859	0.784	0.909	0.842	0.889	0.625	0.734	0.571	0.750	0.649
EPIR	**0.854**	**0.899**	**0.876**	**0.849**	**0.903**	**0.875**	**0.817**	**0.909**	**0.861**	**0.889**	**0.662**	**0.759**	**0.726**	**0.750**	**0.738**

ing benefits beyond fine-tuning.

We also conducted the experimental results with the standard deviation over multiple runs. Specifically, on the ICEWS-18 dataset, the average standard deviations of Precision, Recall, and F1 are 0.026, 0.037, and 0.016, respectively. The relatively small deviations indicate that our method achieves stable performance across different runs.



(a) EPIR components effectiveness (ICEWS-15) (b) Varying s (ICEWS-15)



(c) Varying l (ICEWS-15) (d) Varying c (ICEWS-16)

Figure 1: Performance Evaluation

Exp-2: Ablation Studies. We conduct ablation studies to examine the contribution of individual components in EPIR, focusing on evidence modeling, structural mechanisms, LLM-based augmentation and environment conditioning.

Evidence-Level. We first assess the necessity of combining observational and structural evidence. Removing observational evidence (EPIR w/o Obs) leads to a significant drop in recall, indicating that the event-chain-based context is crucial for identifying potential influences. In contrast, removing structural mechanisms entirely (EPIR w/o Struct) causes a substantial decline in precision, as predictions become dominated by local co-occurrence patterns.

LLM-Augmented Mechanism. To evaluate the effect of LLM-based mechanism augmentation, we compare EPIR with a variant that uses only data-mined core mechanisms. Performance consistently improves after incorporating LLM-augmented mechanisms, particularly on sparse datasets, confirming that LLM generation effectively extends mechanism coverage beyond what can be reliably mined from data alone.

Environment Conditioning. We analyze the impact of environmental context by selectively removing individual environment dimensions, including temporal, spatial, and semantic attributes, from mechanism activation. As shown in Table 2, disabling any single factor consistently degrades performance, with the most drops observed in recall.

Training Objective Ablation. We further examine the effect of contrastive learning by removing the contrastive objective while keeping all other components unchanged. As shown in Figures 1(a), Without contrastive learning, performance consistently degrades across all datasets, with a more

pronounced drop in recall.

Heavy Reliance on Heuristic Chain Construction. To assess the impact of heuristic chain construction on EPIR, we injected three types of noise into the event chains: event replacement, event deletion, and event swapping. The results on ICEWS16 are reported in Table 3. The performance degrades gradually rather than collapsing abruptly. Under moderate noise levels (10%–20%), both Precision and Recall remain above 0.6, and a substantial decline appears only when the noise level reaches 30%. These results indicate that EPIR is sensitive to severe structural corruption, but does not rely on heuristic chain construction.

Table 3: Impact of injected noise ratio on EPIR performance on ICEWS-16.

Ratio	Precision	Recall	F1
0.05	0.762	0.734	0.748
0.10	0.758	0.642	0.695
0.15	0.732	0.636	0.680
0.20	0.689	0.619	0.652
0.30	0.607	0.483	0.538

Exp-3: Parameter Sensitivity. We analyze the sensitivity of EPIR to key hyperparameters, including (1) negative sample size s, (2) event chains length l, and (3) the confidence threshold c. All other parameters are fixed to their default settings.

Varying s. As shown in Figures 1(b), increasing s generally improves recall while slightly reducing precision. This shows that: more negative samples push the model to better differentiate between positive and negative pairs. However, these samples also raise the risk of false positives.

Varying l. We analyze the effect of event-chain length l by varying the maximum contextual span used for mechanism activation. As shown in Figures 1(c), increasing l leads to gradual declines in precision, recall, and F1-score. This indicates that excessively long chains introduce additional contextual noise, weakening the reliability of activated structural signals. Nevertheless, the performance degradation remains moderate, suggesting that EPIR is robust to limited over-extension of contextual scope, and that its multi-hop mechanisms are primarily driven by informative intermediate structures rather than an unbounded chain length.

Varying c. As shown in Figures 1(d), lowering c generally increases recall at the cost of a slight drop in precision, reflecting a trade-off between schema coverage and noise, where c is a frequency-normalized signal for schema candidate selection.

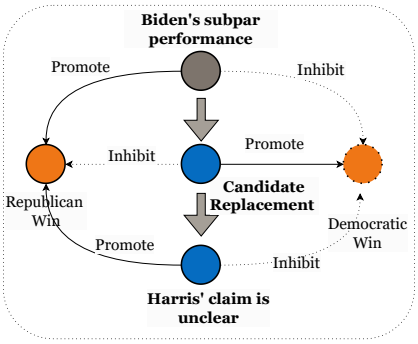


Figure 2: Case Study

Exp-4: Case Study. Using election-related events from the 2024 U.S. election corpus (Wu et al., 2025), EPIR finds that campaign speeches with a negative public reception promote subsequent discussions on candidate replacement, while an ambiguous policy communication promotes shifts in voter sentiment toward opposing candidates, consistent with observed dynamics in the 2024 U.S. election.

Error Analysis. We analyze the types of event pairs that EPIR misclassifies and find that the errors mainly arise from ambiguous transitions, missing context, short temporal gaps, and fine-grained distinctions between similar event types. Specifically, most errors fall into the following situations: (1) branch ambiguity, where the same event prefix can reasonably lead to different next events (e.g., Make statement → Consult → Engage in negotiation may either continue verbally or escalate to Threaten in GDELT); (2) missing context, such as absent text or location information, which makes it harder to distinguish actors or countries; (3) short time gaps (0–3 days), where closely timed but logically independent events are grouped into the same chain; and (4) fine-grained confusion between semantically similar diplomatic categories (e.g., Make statement vs. Make an appeal, Consult vs. Express intent to cooperate in ICEWS), particularly under minor actor or timing changes.

6 Conclusion

We presented **EPIR**, a framework for analyzing promoting and inhibiting relationships. EPIR formalizes directional influence as a relative effect and decomposes it into explicit evidence and latent mechanisms. Experiments demonstrate that EPIR outperforms existing approaches. Future work includes developing a practical system.

Limitations

Despite its effectiveness, EPIR has several limitations. First, the quality of promoting–inhibiting inference is inherently dependent on the completeness and reliability of the constructed event context. Missing or noisy contextual attributes can weaken the model's ability to distinguish genuine directional influence from spurious correlations. In such cases, both the observational scoring module and the mechanism-based reasoning component may produce less reliable estimates. Second, EPIR imposes implicit requirements on the quality of event-chain construction. Since directional influence is inferred relative to a historical event chain, errors in event extraction can propagate through the model and distort contextual alignment and mechanism activation. In particular, fragmented or overly short event chains may fail to capture the long-range dependencies necessary for modeling indirect or multi-hop influence.

Acknowledgments

We thank the anonymous reviewers for their valuable suggestions to improve the quality of this work. This work was supported by the General Program of the National Natural Science Foundation of China (62572345) and the Tianjin Natural Science Foundation (24JCQNJC00810).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger. 2009. Abnormal events detection based on spatio-temporal co-occurences. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2458–2465. IEEE.

Debarun Bhattacharjya, Tian Gao, Nicholas Mattei, and Dharmashankar Subramanian. 2021. Cause-effect association between event pairs in event datasets. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1202–1208.

Debarun Bhattacharjya, Karthikeyan Shanmugam, Tian Gao, Nicholas Mattei, Kush Varshney, and Dharmashankar Subramanian. 2020. Event-driven continuous time bayesian networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3259–3266.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. Icews coded event data. *Harvard Dataverse*, 12.

Yi Cai, Qing Li, Haoran Xie, Tao Wang, and Huaqing Min. 2013. Event relationship analysis for temporal event search. In *Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, Wuhan, China, April 22-25, 2013. Proceedings, Part II 18*, pages 179–193. Springer.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. 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)*, pages 4862–4872.

Jing Chen, Xinyu Yang, Mingxin Liu, and Miaomiao Liu. 2024. Sdegnn: Signed graph neural network for link sign prediction enhanced by signed distance encoding. *The Journal of Supercomputing*, pages 1–25.

Joscha Cüppers, Sascha Xu, Musa Ahmed, and Jilles Vreeken. 2024. Causal discovery from event sequences by local cause-effect attribution. *Advances in Neural Information Processing Systems*, 37.

Tyler Derr, Yao Ma, and Jiliang Tang. 2018. Signed graph convolutional networks. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 929–934. IEEE.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687.

Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. 2025. Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 543–554.

Zhiyu Fang, Shuai-Long Lei, Xiaobin Zhu, Chun Yang, Shi-Xue Zhang, Xu-Cheng Yin, and Jingyan Qin. 2024. Transformer-based reasoning for learning evolutionary chain of events on temporal knowledge graph. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 70–79.

Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma,

Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Zhu Hua, Huang Hong, Yin Kehan, Fan Zejun, Jin Hai, and Liu Bang. 2024. Causalnet: Unveiling causal structures on event sequences by topology-informed causal attention. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*.

Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. 2019. Signed graph attention networks. In *International Conference on Artificial Neural Networks*, pages 566–577. Springer.

Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. 2021. Sdgnn: Learning node representation for signed directed networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 196–203.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. *arXiv preprint arXiv:2305.17607*.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global database of events, language, and tone. In *ISA Annual Convention*, page 0.

Yanhao Li and Wei Liu. 2022. Sudden event prediction based on event knowledge graph. *Applied Sciences*, 12(21):11195.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth international conference on the principles of knowledge representation and reasoning*.

Feiteng Mu and Wenjie Li. 2023. Enhancing event causality identification with counterfactual reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 967–975.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Meng Qu, Junkun Chen, Louis-Pascal Xhonneux, Yoshua Bengio, and Jian Tang. 2021. Rnnlogic: Learning logic rules for reasoning on knowledge graphs. In *International Conference on Learning Representations*.

Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Yongqiang Zhao, Fang Wang, and Chongyang Tao. 2023. Seag: Structure-aware event causality generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4631–4644.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706.

Ting-Ting Wu, Xiao Ding, Li Du, Bing Qin, and Ting Liu. 2024. Reasoning subevent relation over heterogeneous event graph. *Knowledge and Information Systems*, pages 1–23.

Weiqi Wu, Shen Huang, Yong Jiang, Pengjun Xie, Fei Huang, and Hai Zhao. 2025. Unfolding the headline: Iterative self-questioning for news retrieval and timeline summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4385–4398.

Pinghua Xu, Wenbin Hu, Jia Wu, and Bo Du. 2019. Link prediction with signed latent factors in signed social networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1046–1054.

A Appendix: Implementation Details and Additional Experimental Results

A.1 Conceptual Clarifications

Comparable contexts. We represent the context at time t as $\mathbf{C}(t) = (\mathbf{C}^{\text{struct}}, \mathbf{C}^{\text{env}}(t), \mathbf{C}^{\text{obs}})$, where $\mathbf{C}^{\text{struct}}$ denotes stable structural conditions, $\mathbf{C}^{\text{env}}(t)$ denotes time-varying environmental states, and $\mathbf{C}^{\text{obs}}$ denotes the observation and measurement setting. Two events are said to be observed under *comparable contexts* if they share the same structural context $\mathbf{C}^{\text{struct}}$ and observational context $\mathbf{C}^{\text{obs}}$. Under comparable contexts, environmental states are allowed to vary over time, but their evolution admits a consistent interpretation and supports meaningful contrast estimation.

Observational Contrast under Comparable Contexts. Throughout this paper, the promoting or inhibiting relationship between two events is defined via an *observational contrast* conditioned on comparable contextual states, rather than an interventional effect.

Specifically, the quantity $P(v_j \mid v_i, \mathbf{C})$ conditions on the co-occurrence of v_i and v_j under similar world contexts $\mathbf{C}$, instead of enforcing an external intervention on v_i. Accordingly, $\neg v_i$ denotes the non-occurrence of v_i under comparable contexts, rather than a counterfactual removal. This distinction is critical for event-centric observational data, where controlled interventions are generally unavailable.

A.2 Implementation Details of EPIR

Figure 3 illustrates the data flow and module alignment used in our implementation. This figure is provided to support implementation and reproducibility, rather than to introduce new modeling components.

Example 2: Consider whether the release of a GPT model promotes Microsoft's subsequent investment in OpenAI. EPIR first builds time-ordered event chains and abstracts them into type-level patterns. The observational module estimates a context-matched contrast under comparable historical and environmental conditions:

$$\begin{aligned}\Delta P = {} & P(\textit{Investment} \mid \textit{ModelRelease}, C) \\ & - P(\textit{Investment} \mid \neg\textit{ModelRelease}, C),\end{aligned}$$

where C denotes the matched context. If $\Delta P > 0$, this provides positive observational evidence.

The structural module further activates a mined mechanism m: *ModelRelease* $\rightarrow$ *UserAdoptionExplosion* $\rightarrow$ *StrategicInvestment*, when it satisfies support and confidence thresholds and preserves temporal order. Finally, EPIR fuses the observational and structural signals to make the prediction. □

A.2.1 Event Chain Construction Notes

Motivation and limitation of existing constructions. Existing event chain construction methods are typically designed for narrative coherence, topical grouping, or heuristic causal ordering. Such constructions are not directly applicable to our setting, where the goal is to estimate promoting or inhibiting relationships via *context-conditional observational contrasts*. In particular, generic chains may freely mix events observed under heterogeneous world conditions, which violates the requirement that the contrast between v_i and v_j be evaluated under comparable contexts. Moreover, these methods do not explicitly support alignment with time-varying environmental signals, which are required by EPIR to model contextual states. For this reason, we implement a lightweight and reproducible event chain constructor tailored to our observational setup. This component is not a modeling contribution of the paper, but is introduced solely to support end-to-end reproducibility.

Pair-conditioned chain construction. Given a target event pair (v_i, v_j) with timestamps (t_i, t_j), we construct an event chain by retrieving candidate events from raw logs within a fixed temporal window $[\min(t_i, t_j) - \Delta,\ \max(t_i, t_j) + \Delta]$, where Δ is a dataset-independent temporal margin. Events are first normalized into a unified schema and indexed by timestamp. To avoid mixing events from incomparable world settings, we apply a deterministic eligibility filter based on a *context regime key*, which captures stable attributes shared by the target pair and the environmental measurements (e.g., location or system identifier, depending on the dataset). Only events matching the same regime key as (v_i, v_j) are retained. The remaining events are sorted by timestamp to form an event chain $\langle v_1, \ldots, v_k \rangle$, where each event is represented as $v = (t, \ell, \mathbf{L}, \mathbf{S})$ following Section 3.2.

Dynamic environments along the chain. Event chains are constructed prior to the integration of environmental signals and are defined purely over

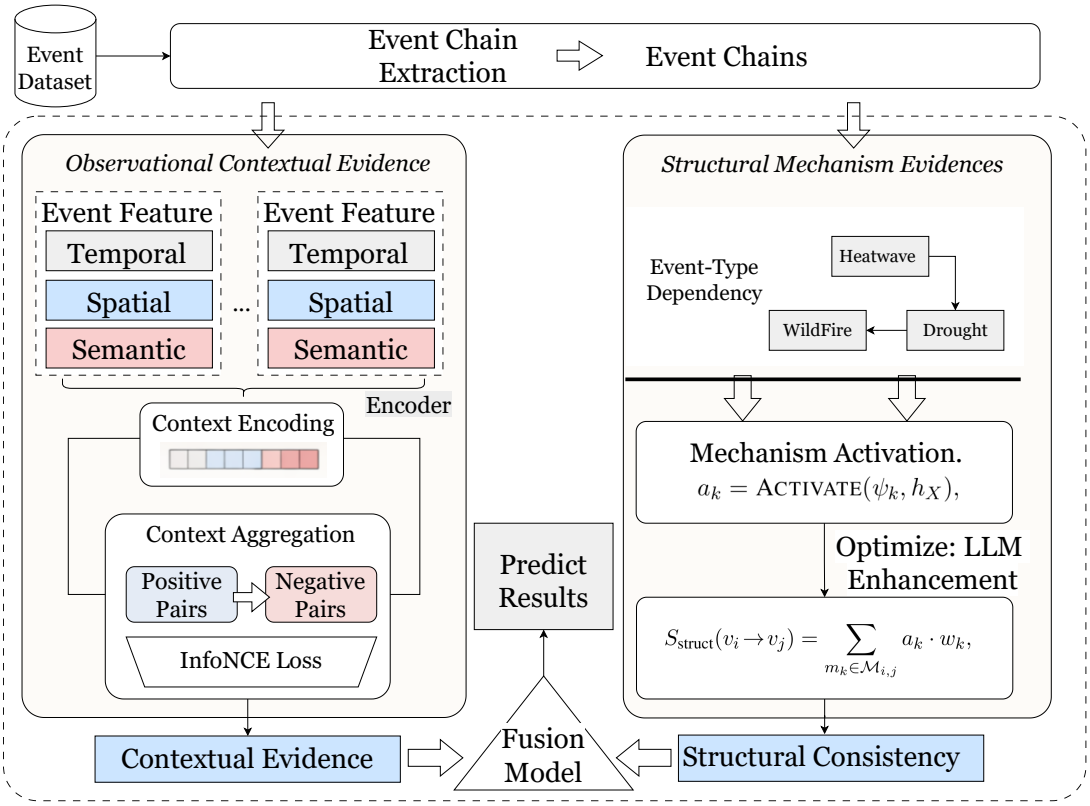


Figure 3: Illustration of EPIR workflow.

events and their temporal order. The environmental context is not assumed to be static. Instead, each event v_r in the chain is later associated with a time-varying environmental state $\mathbf{C}(t_r)$ through temporal alignment. The comparability requirement in chain construction enforces a shared environmental *regime*, rather than a fixed environmental state, ensuring that the dynamic evolution of $\mathbf{C}(t)$ along the chain is meaningful and comparable. Environmental information is therefore introduced after chain construction to parameterize contextual states, rather than to determine which events are included in the chain.

A.2.2 Contextual Encoder for Event Chains (Implementation Detail)

This component implements the *contextual encoder* described in Section 4, which maps an event chain observed under comparable contexts to contextualized event representations used for directional contrast estimation. It does not introduce a new modeling assumption and serves solely as an implementation detail of EPIR.

Given an event chain $\langle v_1, \ldots, v_k \rangle$ constructed as in Appendix A.2.1 and the corresponding time-varying environmental states $\{\mathbf{C}(t_1), \ldots, \mathbf{C}(t_k)\}$ obtained in Appendix A.2.3, each event is represented by an embedding that combines event attributes and aligned environmental features. The encoder aggregates these representations along the chain to capture both local event-level interactions and longer-range contextual dependencies, producing contextualized embeddings for subsequent contrast estimation.

Mixing Unit Architecture. Motivated by (Fang et al., 2024), Figure 4 illustrates the internal structure of a single mixing unit used in the contextual encoder. The input to the mixing unit is an event-chain representation $\mathbf{E} \in \mathcal{R}^{k \times d}$, where k denotes the number of events in the chain and d is the embedding dimension of each event. The output has the same shape and serves as a refined contextual representation that incorporates both intra-event and inter-event dependencies.

The mixing unit adopts a pre-normalization design and consists of two sequential stages. First, the input $\mathbf{E}$ is normalized using layer normalization to stabilize feature distributions, followed by a transposition operation that rearranges tensor dimensions to enable feature-wise processing. A channel-wise MLP is then applied, operating inde-

pendently on each event while mixing information across feature channels. This stage captures interactions among temporal, spatial, and semantic features within the same event, enriching the intra-event representation.

The output of the channel mixing stage is again normalized and transposed, switching the processing axis from feature-wise to event-wise. A patch-wise MLP is subsequently applied across the event dimension, allowing information to propagate among different events in the chain. This stage models long-range dependencies and latent interactions between temporally distant but semantically related events, which are essential for capturing indirect influence propagation.

Finally, the output of the patch-wise MLP is mapped back to the original tensor layout, producing an output representation $\mathbf{E}' \in \mathcal{R}^{k \times d}$ with the same dimensionality as the input. By alternating channel-wise and patch-wise mixing within a single unit, the model jointly captures fine-grained feature interactions and global event-level context, enabling efficient contextual aggregation without explicit attention mechanisms.

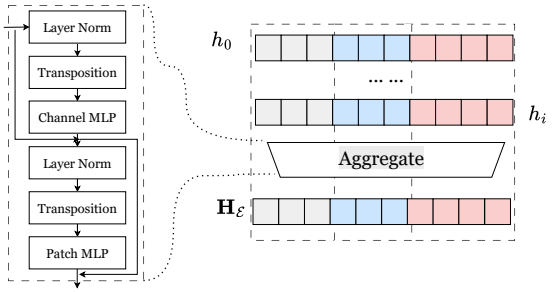


Figure 4: Illustration of Mixed Module.

Environmental Signal Alignment. Given an event chain constructed as described in Appendix A.2.1, environmental signals are aligned to individual events based on temporal correspondence. For each event timestamp t_r, environmental observations within a predefined temporal neighborhoods are aggregated and encoded as contextual features associated with the event. This procedure yields a time-varying environmental state $\mathbf{C}(t_r)$ for each event in the chain, without imposing any assumption that environmental conditions are static or identical across events. Environmental alignment serves solely to parameterize contextual states used by EPIR, and does not affect event chain construction or the definition of comparable contexts.

Negative Sampling Strategy. Negative samples are generated by constructing event pairs that preserve contextual similarity while breaking the temporal or semantic dependency between events. Specifically, we apply: (i) temporal permutation within comparable windows, (ii) cross-location substitution under similar environments, and (iii) semantic replacement using type-consistent but unrelated events. This strategy ensures meaningful contrastive supervision.

(1) Temporal Shuffling (TS). *TS-easy* samples events n whose timestamps violate temporal consistency, i.e., $t(n) \leq t(v_i)$ or $t(n) - t(v_i) > \Delta t_{\text{neg}}$ (with $\Delta t_{\text{neg}} = 60$ days). *TS-hard* samples events from $[t(v_i),\, t(v_i) + \Delta t_{\text{pos}}]$ such that $\text{type}(v_i) \rightarrow \text{type}(n)$ is an infrequent corpus-level transition.

(2) Spatial Transposition (SP). *SP-easy* replaces the positive event with another event occurring beyond a spatial radius $R_{\text{neg}} = 500$ km, while keeping the temporal window unchanged. *SP-hard* preserves the country (or first-level administrative region) but shifts the event to a different city cluster, maintaining temporal proximity while altering the local spatial context.

(3) Semantic Mutation (SM). Semantic mutation modifies event semantics while preserving structural validity, including role arity and required argument slots: (i) *Entity swap*, replacing a subject or object with another entity of the same role type whose embedding similarity lies within $[\alpha_{\min}, \alpha_{\max}] = [0.6, 0.9]$, avoiding co-reference; (ii) *Type perturbation*, replacing the event type with an ontological or a co-occurrence neighbor that is semantically similar but not a frequent successor; (iii) *Argument mutation*, modifying optional arguments via synonym substitution or masking, while keeping required roles intact.

Validity Filtering and Hard-Negative Mining. To prevent noisy or false negatives, a candidate negative n is discarded if (1) its contextual similarity satisfies $s(z, n) \geq m$ (with $m = 0.6$), measured using the current encoder or a frozen teacher model; or (2) its type transition ranks within the top $\epsilon = 5\%$ of global successor likelihood. From the remaining candidates, hard-negative mining selects the top-k nearest negatives to the positive embedding.

A.3 Experimental Reproducibility Details

We next propose our experimental setting and other experimental results in detail.

Datasets. We evaluate EPIR on five real-world event datasets that span diverse temporal scopes, structural properties, and levels of contextual completeness, including (1) **ICEWS-15**, **ICEWS-16**, and **ICEWS-18** (Boschee et al., 2015) are large-scale, fully structured political and social event datasets, covering different years (2015, 2016 and 2018) with consistent annotation schemas. Each dataset provides fine-grained temporal, spatial, actor, and CAMEO-type information; (2) **GDELT** (Leetaru and Schrodt, 2013) offers broader temporal coverage and global diversity, augmented with sentiment and tone indicators derived from news text. We sample 100,000 events from the 1979–2014 period to assess EPIR's robustness in large-scale, noisy, and weakly structured environments; and (3) **OPEN-TLS** (Wu et al., 2025) captures semantically coherent event chains constructed from raw news articles. Unlike ICEWS and GDELT, it lacks explicit geographic annotations and provides a limited structured context.

Annotation Protocol. We construct labels for promoting and inhibiting relationships using a human-centered annotation protocol with LLM-assisted pre-screening. Candidate event pairs are extracted from event chains within a fixed temporal horizon. For each pair, both a frozen large language model and human annotators are provided with the same contextual information, including the event pair, its surrounding event chain, and relevant environmental factors (e.g., time, location, and policy or infrastructure conditions). The LLM is used in an inference-only setting to generate preliminary judgments and rationales, which serve as auxiliary signals. Three domain experts then independently assign labels (*promote*, *inhibit*, or *neutral*); we retain only pairs with agreement from at least two experts and discard cases with no confident consensus. The resulting labels are used as ground truth for model training and evaluation.

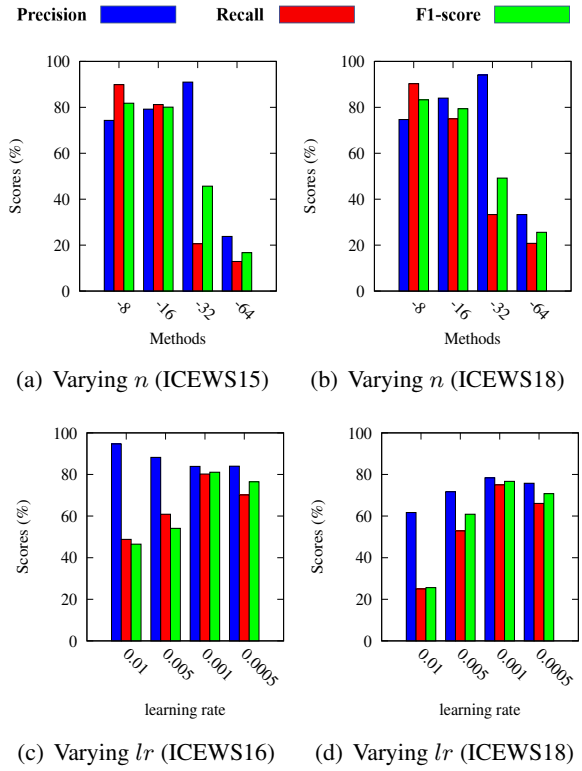
Baselines. We compare EPIR with representative baselines covering multiple paradigms for event relationship analysis, grouped as follows. (1) **Graph-based methods**, which model event relationships using signed or directed graphs and learn node representations via message passing, including SLF (Xu et al., 2019), SGCN (Derr et al., 2018), SIGAT (Huang et al., 2019), SDGNN (Huang et al., 2021), and SDEGNN (Chen et al., 2024). These methods are effective at capturing local structural dependencies but do not explicitly model context-dependent or multi-hop influence mechanisms. (2) **Sequential rule-based reasoning methods**, represented by RNNLogic (Qu et al., 2021), which performs neural rule induction over event sequences and provides interpretable path-based reasoning. (3) **Causal analysis approaches**, including Cascade and causalNet (Cüppers et al., 2024; Hua et al., 2024), as well as the LLM-based causal discovery method LLM-cd (Du et al., 2025). These methods serve as references for modeling directional influence under explicit causal assumptions, but are not specifically designed for context-dependent promoting or inhibiting relationships from observational data. (4) **LLM-based approaches**, including GPT-4 and GPT-4o (Achiam et al., 2023), as well as DeepSeek-v3 (Liu et al., 2024) and DeepSeek-R1 (Guo et al., 2025), which are evaluated in a prompt-based inference setting for event relationship reasoning.

Baselines Approaches. For baselines with publicly available runnable code, we use the official implementations with default hyperparameters, requiring only minimal data-format adaptations to ensure compatibility for our event datasets. For methods without usable released code, we re-implement them based on the original papers, with only minimal task-specific adjustments when necessary.

Graph-based baselines are applied by constructing event interaction graphs, where nodes represent event instances, and edges encode promoting and inhibiting relationships within event chains. For the sequential rule-based method RNNLogic, which is originally evaluated using ranking metrics (e.g., Hits@K, MRR), we reinterpret its output scores as confidence values and map them to binary promoting/inhibiting predictions via thresholding, enabling evaluation using standard classification metrics (Precision, Recall, and F1). Causal analysis methods are included as reference baselines for modeling directional influence. Since our task focuses on observational promoting–inhibiting relationships rather than interventional causal effects, we directly use their inferred influence or causality scores as directional signals and evaluate them under the same classification protocol. For LLM-based baselines, we adopt a unified prompt template and structured output format across all models, using inference-only settings without fine-tuning. For local LLM-based baselines, we evaluate both the original locally deployed models and

their supervised fine-tuned variants trained on each dataset's training split.

Supplementary Parameters Setting. For EPIR, we set the batch size to 16 and the learning rate to 0.05. The temperature for the contrastive loss is fixed at 0.11. We use an embedding dimension of 8, with a network depth of two layers. The attention dropout rate is set to 0.5. We adopt a maximum rule length of 4 and set the rule probability threshold to 0.001, unless stated otherwise.



(a) Varying n (ICEWS15) (b) Varying n (ICEWS18)

(c) Varying lr (ICEWS16) (d) Varying lr (ICEWS18)

Figure 5: Performance Evaluation

Parameters Sensitivity. We explore the supplementary performance caused by other parameters.

(1) Varying n. We evaluate the effect of event attribute dimension over *ICEWS* datasets. We find the optimal n(i.e., 8, 16, 32, 64) for the framework by setting the parameters to their default values. As shown in Figure 5(a) to 5(b), when n increases from 8 to 32, recall decreases while precision increases. This aligns with the expectation that as the dimensionality of the event attribute embedding increases, it becomes easier to distinguish between data points, thereby increasing precision. However, the distribution of data points becomes sparser, leading to a decrease in recall. We also observe a dimensionality catastrophe when n is 64.

(2) Varying lr. We evaluate the impact of the learning rate in *MCSR*. We vary the learning rate (i.e., 0.01, 0.005, 0.001, 0.0005) and then analyze the effect on hidden association learning. As shown in Figures 5(c) to 5(d), we demonstrate that the model performs best when the learning rate is set to 0.001.

Additional Experiment on Event Chain Quality with OpenTLS. We conduct an additional experiment on OpenTLS using a degraded variant, OpenTLS-L, where temporal and entity associations in event chains are partially corrupted or omitted. As shown in Table 4, EPIR drops from 0.649 to 0.375 in F1, with corresponding declines in Precision and Recall. This result confirms that higher-quality event chains lead to more reliable reasoning.

Table 4: EPIR performance comparison on original and degraded event chains (OpenTLS vs. OpenTLS-L).

Dataset	Precision	Recall	F1 Score
OpenTLS-L	0.375	0.375	0.375
OpenTLS	**0.571**	**0.750**	**0.649**

Discussion. We discuss the balance between complexity and interpretability. EPIR's complexity stems from its modular design rather than a black-box structure. Each module is independent: the structural module outputs explicit mechanisms and paths, the observational module computes environment-matched contrasts, and the fusion layer combines them in weighted form. For each prediction, EPIR reports the activated type-level mechanisms and supporting paths, enabling traceability to specific rules (as shown in the error analysis). Ablation results further confirm that structural and observational components contribute independently.