Suggest-Verify-Revise: A Three-Stage Document-Level Event Causality Identification with Narrative Consistency

Ya Su[1], Hu Zhang[1,2]*, Dan Qiao[1], Yujie Wang[1], Yunxiao Zhao[1], Yue Fan[1], Shike Li[1,2], Ru Li[1,2], Hongye Tan[1,2]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
su_ya6990@163.com, {zhanghu,qiaodan,lsk,liru,tanhongye}@sxu.edu.cn, init_wang@foxmail.com, {yunxiaomr,yuefan24}@163.com

Abstract

Document-level Event Causality Identification (DECI) aims to identify causal relations among multiple events within unstructured text. Existing methods predominantly rely on local semantic similarity for independent event-pair discrimination, thereby overlooking the influence of the overall narrative backbone in the propagation of causal dependencies and the role differentiation of events within multi-cause/multi-effect structures. Therefore, we propose a Suggest-Verify-Revise approach for document-level Event Causality Identification with narrative consistency (SVRECI). In the suggest stage, we integrate multi-dimensional heuristic causal suggestions generated by an LLM with structural suggestions derived from hypergraph modeling to provide multi-source initial support for candidate event pairs. In the verify stage, we introduce a Topological Hawkes process to perform constrained verification of narrative propagation consistency among events. In the revise stage, we construct a dynamically evolving document-level causal graph and incorporate a structure-aware dual-level contrastive learning mechanism at both the event and event-pair levels, iteratively reducing noisy edges over multiple iterations. Experimental results on EventStoryLine and Causal-TimeBank datasets demonstrate that our approach outperforms previous methods.

1 Introduction

Reasoning about the causal dependencies between events is a core cognitive ability in human narrative comprehension. Document-level Event Causality Identification (DECI), a critical task in Natural Language Processing (NLP), aims to determine whether a causal dependency exists between any two events within unstructured text. This task is widely applied in scenarios such as question answering (Oh et al., 2017; Liu et al., 2023b; Xu et al., 2025), narrative generation (Ammanabrolu et al., 2021), event prediction (Radinsky et al., 2012), and sentiment analysis (Zhao et al., 2026).

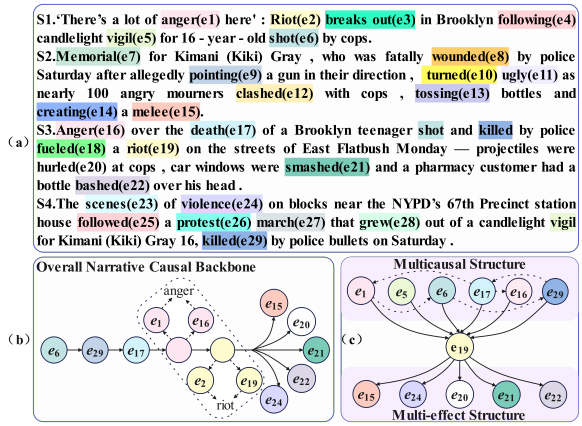

Figure 1: (a) A dataset example with 4 sentences and 29 event mentions (same colors indicate semantically similar mentions); (b) the document-level narrative causal backbone, where dashed lines denote conceptual aggregation; and (c) the multi-cause and multi-effect structure, with dashed lines indicating potential causal relations among multiple causes.

Unlike sentence-level causal judgment, DECI requires modeling cross-sentence long-distance dependencies, multi-event combinatorial structures, and cascading propagation relations. These characteristics introduce higher structural complexity and reasoning uncertainty into causal modeling. Existing methods (Wu et al., 2023; Phu and Nguyen, 2021; Liu et al., 2023a) generally treat event pairs as independent samples for local semantic discrimination, overlooking the overall narrative causal backbone formed within a document. This makes it difficult to characterize the role differentiation and concept-level aggregation of events within multi-cause and multi-effect structures. In coherent narratives, causal relations do not exist in isolation, but instead form a hierarchically organized storyline that runs through the entire discourse (Trabasso

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19067–19091
July 2-7, 2026

et al., 1982; Sah, 2013), namely the narrative causal backbone. Within it, causes precede their effects, and multiple causal links are temporally sequenced, forming narrative consistency–that is, the orderly unfolding of causal relations over time.

We focus on two major challenges in DECI: **(1) overlooking the role of the narrative causal backbone.** In real-world documents, causal relations are not composed of discrete event pairs, but are instead organized around a few structurally central events into a propagating narrative causal backbone. As shown in Figure 1(b), even four sentences can form a backbone (e.g., $shot_{e_6} \rightarrow killed_{e_{29}} \rightarrow death_{e_{17}} \rightarrow anger_{e_1} \rightarrow riot_{e_2}$), where central events e_6, e_{29}, and e_2 appear multiple times. Moreover, multiple similar mentions may refer to the same conceptual event, as shown in Figure 1(a), such as "anger/Anger", "riot/Riot", "followed/following", and "killed/killing". Without narrative consistency constraints and concept-level aggregation, the overall causal structure becomes fragmented. **(2) Insufficient modeling of multi-cause and multi-effect structures and causal roles.** In real-world documents, causal relations are not determined by independent event pairs; instead, events generally participate in multi-cause and multi-effect shared structures. A single event may serve as a common cause for multiple subsequent events and may also be triggered by multiple prior causes. Moreover, event e_{19} can trigger e_{20}, e_{21}, and e_{22}, while e_6, e_{29}, e_{17}, and e_{16} jointly trigger e_{19}, as shown in Figure 1(c). Ignoring such role differentiation makes it difficult to distinguish the asymmetric cause-like and effect-like roles of events in complex causal structures.

To address the above challenges, we propose SVRECI, an approach that models the DECI task as a dynamic evolutionary process within a unified narrative causal space. In the suggest stage, we integrate multi-dimensional heuristic causal suggestions generated by an LLM with structural suggestions derived from hypergraph modeling to provide multi-source initial support for candidate event pairs. In the verify stage, we introduce a topological Hawkes consistency verification mechanism to iteratively validate candidate causal edges from the perspective of narrative propagation consistency, gradually forming a globally self-consistent narrative causal backbone. In the revise stage, we construct a dynamically evolving document-level directed causal graph, combined with a structure-aware dual-level contrastive learning mechanism at both the event and event-pair levels. This enhances the causal role differentiation of events and iteratively re-evaluates or suppresses low-consistency causal edges during the verify-revise cycles. The main contributions of this work are summarized as follows:

- We propose a Suggest-Verify-Revise approach that formulates document-level event causality identification as a dynamic evolutionary process within a unified narrative causal space.
- We introduce a topological Hawkes-based consistency verification mechanism to globally examine the narrative propagation plausibility of candidate causal edges and perform iterative correction.
- Experimental results on two public benchmark datasets demonstrate that SVRECI significantly outperforms existing methods.

2 Related Work

Early DECI studies relied on lexical, syntactic, and statistical features to construct classifiers (Beamer and Girju, 2009; Gao et al., 2019). With the development of deep learning, research has gradually shifted to four major paradigms: **Knowledge Augmentation** (Wu et al., 2023; Ding et al., 2024; Su et al., 2025b), which inject external knowledge but is limited by knowledge coverage and rule quality; **Graph-based Reasoning** (Phu and Nguyen, 2021; Chen et al., 2022; Liu et al., 2024), which models multi-hop event dependencies via global document structures but is sensitive to noisy graph edges; **Prompt Learning** (Liu et al., 2023a; Xiang et al., 2025b), which reformulates ECI as generation or completion tasks but heavily depends on manual prompt design and lacks structural constraints; and **Contrastive Learning** (Ding et al., 2024; Su et al., 2025b), which enhances event-pair discrimination via positive and negative samples but requires high-quality sample construction. Overall, DECI research has evolved from shallow feature-based modeling toward deep semantic and structural reasoning. Among these paradigms, graph-based approaches have become the dominant pathway for document-level ECI due to their global structural awareness and multi-hop dependency modeling capability. However, existing methods still generally lack narrative consistency modeling and iterative

verification-revision mechanisms for causal relations.

3 Methodology

Given a document containing an event set $\mathcal{E} = \{e_1, \ldots, e_n\}$, as shown in Figure 1, the DECI model aims to determine whether a causal relation exists between any event pair (e_i, e_j) with $i \neq j$. Figure 2 presents the framework of the proposed *SVRECI* approach, which primarily consists of three modules: (a) causal suggestions (Section 3.1); (b) narrative consistency verification (Section 3.2); and (c) dynamic causal graph revision (Section 3.3). Each module is described in detail below.

3.1 Causal Suggestion

LLM-based Multi-dimensional Heuristic Causal Suggestions. As shown in Figure 2(a), given a candidate event pair (e_i, e_j) and its context, we leverage the DeepSeek LLM (Liu et al., 2025) to generate multi-perspective chain-of-thought rationales and map them into confidence scores in the $[0, 1]$ range, corresponding to semantic, causal-role, dependency, and temporal perspectives, respectively, which are defined as:

$$\mathbf{h}(e_i, e_j) = \left[h_{ij}^{\text{sem}}, h_{ij}^{\text{cau}}, h_{ij}^{\text{dep}}, h_{ij}^{\text{temp}}\right] \quad (1)$$

The detailed heuristics and prompt templates are provided in Appendix B. The overall LLM-based heuristic causal strength is computed as $s_{ij}^{\text{LLM}} = \frac{1}{K}\sum_{k=1}^{K} h_{ij}^{k}$. This score serves only as a soft causal prior to guide subsequent verification and correction modules, rather than as a final causal decision.

Hypergraph-based Structural Causal Suggestions. In narrative documents, events are embedded in shared structural contexts rather than appearing as independent pairs. To avoid introducing premature discrete causal decisions that may amplify structural noise, we model the set of document-level events as a structural hypergraph $g_s = (\mathcal{V}, \mathcal{E}_s)$ without any causal labels or directional assumptions. Here, $\mathcal{V} = \{e_1, e_2, \ldots, e_N\}$ denotes the set of event nodes, and $\mathcal{E}_s$ consists of hyperedges encoding only structural sharing and equivalence constraints, without any explicit causal information. Following the LKCER method (Su et al., 2025b), which aggregates similar event mentions into conceptual events, we further incorporate event coreference to construct hyperedges. For each coreference cluster $C_k \in \mathcal{C}$, a coreference hyperedge $h_k^{\text{coref}} = C_k$ is created to capture the structural equivalence among multiple event mentions referring to the same conceptual event. For any event pair (e_i, e_j), we extract a multi-dimensional structural support vector f_{ij}^{struct} from the hypergraph (see Appendix C for details). This vector comprises three components: it captures the binding strength in shared coreference structures, structural positional divergence, and global structural centrality. The first component $f_{ij}^{(1)}$ measures the binding strength of the event pair within shared coreference structures:

$$f_{ij}^{(1)} = \frac{1}{|\mathcal{E}_s|}\sum_{h \in \mathcal{E}_s} \mathbb{I}(e_i \in h \wedge e_j \in h) \quad (2)$$

The second component $f_{ij}^{(2)}$ represents structural positional divergence, capturing functional differences in narrative roles rather than simple linear textual distance. Each event e is mapped to a low-dimensional positional encoding vector $\phi(e) = \left[\phi_{\text{sent}}(e), \phi_{\text{para-proxy}}(e)\right]$, where $\phi_{\text{sent}}(e)$ denotes the normalized intra-sentence relative position of the event, and $\phi_{\text{para-proxy}}(e)$ encodes the normalized paragraph proxy. Let $\mathcal{R} = \{\text{sent}, \text{para-proxy}\}$ denote the set of positional dimensions. The structural divergence of an event pair (e_i, e_j) is then defined as:

$$f_{ij}^{(2)} = \frac{1}{|\mathcal{R}|}\sum_{r \in \mathcal{R}} |\phi_r(e_i) - \phi_r(e_j)| \quad (3)$$

The third component $f_{ij}^{(3)}$ denotes the global structural centrality, defined as the average hypergraph degree of the two events in the structural hypergraph, reflecting the number of coreference hyperedges each event belongs to:

$$f_{ij}^{(3)} = \frac{\text{Centrality}(e_i) + \text{Centrality}(e_j)}{2} \quad (4)$$

where $\text{Centrality}(e) = |\{h \in \varepsilon_s \mid e \in h\}|$. Events with higher centrality are more likely to lie at the intersection of multiple latent causal chains. Finally, these multi-dimensional features are aggregated to derive the structural causal strength s_{ij}^{struct}, serving as a soft causal prior at the structure level.

3.2 Narrative Consistency Verification

Causal relationships in real-world narratives are often governed by global constraints, such as narrative order, event role distribution, and overall document structure (see Appendix D). Therefore, we

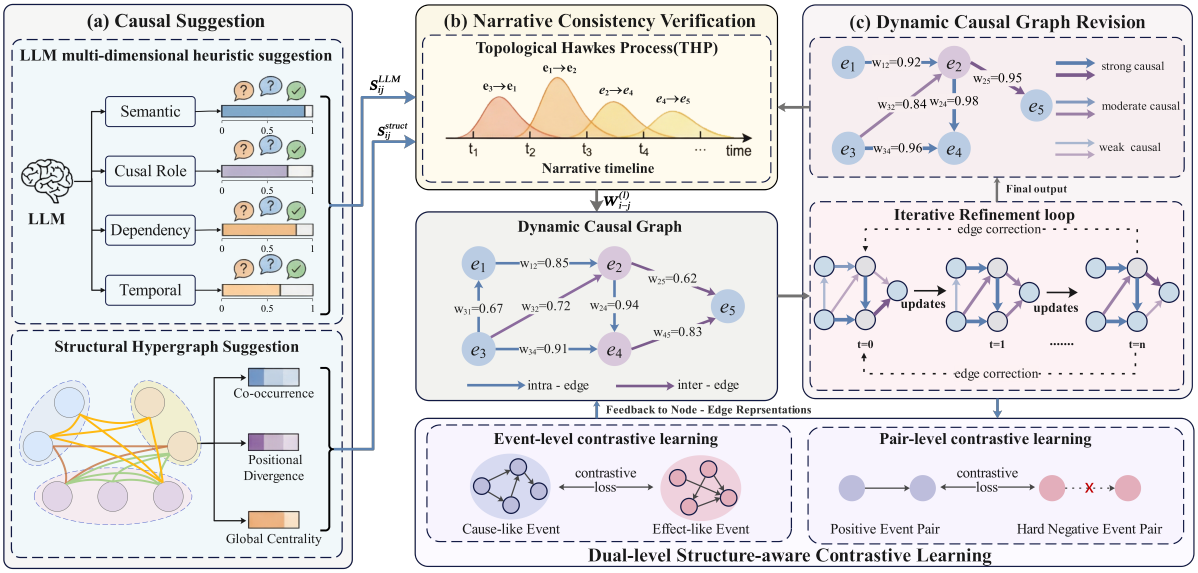

Figure 2: The framework of the proposed *SVRECI* approach, which primarily consists of three modules: (a) causal suggestions; (b) narrative consistency verification via topological hawkes process; and (c) dynamic causal graph revision with dual-level structure-aware contrastive learning.

introduce the Topological Hawkes Process (THP) (Cai et al., 2021; Xu et al., 2016) as a narrative consistency verification module, modeling each document as an ordered sequence of events and enforcing global consistency constraints on the candidate causal structures.

We assign a discrete narrative time index t_i to each event e_i: if a Temporal Link(TLINK) annotation exists, we adopt its relative temporal relationship; otherwise, we use the sentence-level order as a proxy. For any event e_j, its conditional intensity at narrative time t_j is defined as:

$$\lambda_j = \mu_j + \sum_{i \to j} \alpha^{(l)}_{i \to j} \cdot \kappa(t_j - t_i) \cdot \mathbb{I}[(i,j) \in A] \quad (5)$$

where μ_j is a fixed narrative prior parameter, ensuring that even if no causal parent is selected, the event remains an independent narrative unit. $\kappa(\cdot)$ is a kernel function that decays with narrative distance. A denotes the current candidate causal structure and is dynamically updated during iterative optimization, ensuring that consistency verification is always performed on the latest structure. The parameter $\alpha^{(l)}_{i \to j}$ represents the document-level explanatory dependency of e_i on e_j at the l-th iteration, determined by both s^{LLM}_{ij} and s^{struct}_{ij}:

$$\alpha^{(l)}_{i \to j} = \beta_1 \sigma(s^{\text{LLM}}_{ij}) + \beta_2 \sigma(s^{\text{struct}}_{ij}) \quad (6)$$

where β_1 and β_2 are learnable parameters adjusted during training to balance the two types of suggestions. The negative log-likelihood of THP is:

$$\mathcal{J}_{\text{THP}} = -\sum_{j=1}^{N} \log \lambda_j(t_j) + \sum_{j=1}^{N} \int_0^{t_j} \lambda_j(t)\, dt \quad (7)$$

By jointly optimizing all $\alpha_{i \to j}$, THP suppresses spurious causal edges that are locally plausible but globally inconsistent with the document narrative. The final edge weight of the document-level causal graph is given by:

$$w^{(l)}_{i \to j} = \mathbb{E}_{\text{THP}}[\alpha^{(l)}_{i \to j} \cdot \kappa(t_j - t_i)] \quad (8)$$

After graph refinement in the discriminative learning stage, the updated causal matrix is re-submitted to the THP module for further consistency verification. The integration of THP into training is described in Section 3.4.

3.3 Dynamic Causal Graph Revision

Dynamic Document-level Causal Graph Construction and Self-Revision Mechanism. To model global causal dependencies among events at the document level, we construct a weighted directed causal graph $g_c = (\mathcal{V}, \mathcal{E}_c, W)$ where $\mathcal{V}$ denotes the event node set, $\mathcal{E}_c$ is the candidate causal edge set, and $W = \{w_{i \to j}\}$ represents edge weights. To prevent the model from over-relying on structural priors in early training, we adopt a progressive knowledge injection strategy to debias

and perturb the initial weights:

$$w_{i\to j}^{(l)} = \lambda_t \cdot w_{i\to j}^{(0)} + (1-\lambda_t)\cdot \epsilon \tag{9}$$

where ϵ denotes random noise and $\lambda_t \in [0,1]$ increases monotonically with training progress, enabling a smooth transition. Unlike static graph construction methods, we introduce a self-iterative dynamic graph refinement mechanism, which alternates among graph structure update, representation propagation, and edge-level correction across multiple iterations (see Appendix E). The update at the l-th iteration is formulated as:

$$g_c^{(l+1)} = \mathcal{R}\left(\mathcal{V}\left(g_c^{(l)}\right), \boldsymbol{\alpha}^{(l)}\right) \tag{10}$$

where $\mathcal{R}(\cdot)$ is a structural revision operator, $\mathcal{V}(\cdot)$ denotes the consistency verification weight matrix, and $\boldsymbol{\alpha}^{(l)} = \{\alpha_{i\to j}^{(l)}\}$ denotes the document-level causal prior matrix. The refined graph is then used in the next verification round.

At the l-th iteration, the model selects high-confidence event pairs to construct the candidate causal edge set $\mathcal{E}_c^{(l)} = \mathcal{E}_{\text{intra}}^{(l)} \cup \mathcal{E}_{\text{inter}}^{(l)}$ and edge-level consistency is re-evaluated based on updated representations, with the score defined as:

$$s_{i\to j}^{(l)} = \sigma\Big(\mathbf{u}^\top \left[\mathbf{h}_i^{(l)} \,\|\, \mathbf{h}_j^{(l)} \,\|\, \left(\mathbf{h}_j^{(l)} - \mathbf{h}_i^{(l)}\right)\right]\Big) \tag{11}$$

The iteration terminates when prediction changes between successive rounds fall below a threshold and a minimum number of iterations is reached, yielding the final causal graph $g_c^{(l^*)}$. Otherwise, the model continues the iterative *identify–verify–revise* process to progressively suppress noisy edges and reinforce globally coherent causal structures.

Dual-level Structure-aware Contrastive Learning. To further enhance causal structural coherence and boundary discrimination, we propose a dual-level iteration-aware contrastive learning strategy. At each iteration, events and event pairs are dynamically sampled according to the current causal graph to construct positive and negative instances, thereby avoiding representation-structure mismatch from static sampling. Unlike prior work (Li et al., 2025; Chao et al., 2024; Su et al., 2025a), our approach introduces structure-aware contrastive learning that does not rely solely on semantic similarity, suitable for scenarios with few positive samples. Implementation details can be found in Appendix F. Structural contrastive objectives are activated only after the causal graph reaches a preliminary convergence stage and are gradually introduced via weight scheduling to mitigate early-stage noise.

Event-level Structural Contrastive Learning. For each anchor event z_i, we implicitly encode its causal role using the 1-hop causal subgraph. Positive samples z_i^+ are selected from role-consistent events, while negative samples include the most structurally similar role-inconsistent hard negative z_i^{hard} and a randomly sampled easy negative z_i^{rand}. We define $\mathcal{S}_i = \{z_i^+\} \cup \{z_i^{\text{hard}-}, z_i^{\text{rand}-}\}$ and adopt a weighted InfoNCE loss:

$$\mathcal{L}_{\text{event}} = -\log \frac{e^{\cos(z_i, z_i^+)/\tau}}{\sum_{s\in\mathcal{S}_i} w(s) e^{\cos(z_i,s)/\tau}} \tag{12}$$

This objective regularizes event representations toward structural role consistency without directly predicting causal relations.

Event-pair-level Structural Contrastive Learning. Each event pair serves as an anchor, where positive samples are causal pairs, and the structurally most similar non-causal pairs are selected as negatives to enhance boundary discrimination. We employ a margin-based ranking loss:

$$\mathcal{L}_{\text{pair}} = \sum_{(i,j)\in P^+} \Big(\gamma - \sum_s s\cdot \cos(z_{ij}, z_{ij}^s)\Big)_+ \tag{13}$$

where $s \in \{+1, -1\}$ indicates positive or negative samples.

3.4 Training Strategy

We model the multi-round classification loss as a propagation-risk-aware structural loss:

$$\mathcal{L}_{\text{cls}} = \sum_{(e_i,e_j)\in D} \sum_{l=1}^{L_{\text{D}}} \frac{1}{l}\, \ell\left(\hat{p}_{ij,c}^{(l)}\right) \tag{14}$$

where $\hat{p}_{ij,c}^{(l)}$ denotes the predicted probability that the event pair (e_i, e_j) belongs to the true class c at the l-th round, and $\ell(\cdot)$ is the α-balanced focal loss (Liu et al., 2024). The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{event}} + \lambda_2 \mathcal{L}_{\text{pair}} \tag{15}$$

where λ_1 is applied with warm-up to stabilize the event representation space, and λ_2 is gradually activated after classifier convergence to constrain the causal boundaries.

In addition, we employ an alternating optimization scheme to coordinate the discriminative model and the THP-based narrative consistency module.

At each iteration, candidate causal structures predicted by the classifier are refined by the THP module based on the objective in Equation 7, and the resulting structures are used to guide subsequent classification and contrastive learning.

4 Experimental Setting

4.1 Datasets and Evaluation Metrics

We evaluate our method on two publicly available ECI benchmark datasets: EventStoryLine v0.9 (ESC) (Caselli and Vossen, 2017) and Causal-TimeBank (CTB) (Mirza et al., 2014). Dataset statistics are summarized in Appendix G. We adopt Precision (P), Recall (R), and F1-score (F1) as evaluation metrics.

4.2 Baselines

We compare our method with representative existing models on ESC and CTB datasets, including RichGCN (Phu and Nguyen, 2021), ERGO (Chen et al., 2022), CHEER (Chen et al., 2023), SENDIR (Yuan et al., 2023), PPAT (Liu et al., 2023c), iLIF (Liu et al., 2024), EHNEM (Xiang et al., 2025a), GCKAN (Ding et al., 2024), and ESupCL (Li et al., 2025). Following prior zero-shot setups (Gao et al., 2023; Liu et al., 2024) (Figure 15), we also include LLaMA-2-7B and LLaMA-3.1-8B (Touvron et al., 2023), Qwen-2.5 (7B and 14B) (Hui et al., 2024), and DeepSeek (Liu et al., 2025) as LLM baselines to assess the performance of general-purpose generative models on the ECI task. Detailed descriptions of these baselines are provided in Appendix I. We use BERT-base-uncased as the base encoder and optimize the model using AdamW (Loshchilov and Hutter, 2017). Detailed hyperparameter settings are provided in Appendix H.

4.3 Main Results

Table 1 presents our experimental results on the ESC and CTB datasets, highlighting performance differences among methods. Overall, as models incorporate richer structural modeling, performance shows a steady improvement. Cross-sentence causal identification remains a major bottleneck, with most models exhibiting a 12–18% drop in F1 compared to intra-sentence settings , reflecting the intrinsic difficulty of modeling long-range causal chains and document-level narrative consistency. In contrast, SVRECI reduces this gap to under 10% drop in F1 and consistently outperforms previous methods across all settings, indicating its superior capability in capturing global narrative propagation patterns. SVRECI shows higher precision but slightly lower recall, indicating that its structure-aware verification favors reliable causal edges at the cost of missing some potential relations. LLMs achieve high recall but low precision, reflecting a tendency to capture broad causal correlations rather than verifiable structural causal relations. This underscores the value of combining LLM suggestions with structural verification and dynamic revision, and further highlight SVRECI's strength in modeling semantic-structural consistency.

4.4 Ablation Study

We conduct ablation studies to evaluate the contribution of each component in SVRECI. Specifically, we remove the dual-level structure-aware contrastive learning module **-CL**, where **-CL**$_{event}$ and **-CL**$_{pair}$ denote the removal of the event-level and event-pair-level contrastive components, respectively; the iterative causal graph refinement module **-IR**; the topological Hawkes verification module **-TV**; and the proposal generation module **-SG**, where **-SG**$_{LLM}$ and **-SG**$_{Hyp}$ denote the omission of LLM-based heuristic proposals and hypergraph structural proposals.

As shown in Table 2, the revision stage contributes most significantly: removing the CL or IR modules leads to substantial F1 drops, highlighting their critical roles in suppressing noisy edges and strengthening structural discriminability, particularly in challenging cross-sentence scenarios with long-range dependencies and complex narrative structures. These results suggest a clear synergistic effect, where CL improves representation-level discriminability while IR enforces structural consistency over the evolving causal graph. Removing the TV module weakens global narrative consistency and impairs the formation of the narrative causal backbone. In addition, removing the SG module further degrades performance, demonstrating that multi-source initial causal suggestions provide important guidance for subsequent verification and revision.

To further analyze the iterative refinement (IR) module, we conduct finer-grained ablations on its three key components: propagation, pruning, and the iterative mechanism. As shown in Table 3, **-prop.**, **-prun.**, and **-iter.** denote the removal of propagation, pruning, and iterative update operations in the IR module, respectively. **C-F1** and **O-F1** refer to the Cross F1 and Intra&Cross F1

Methods	EventStoryLine									Cause-TimeBank		
	Intra			Cross			Intra and Cross			Intra		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
RichGCN(Phu and Nguyen, 2021)	49.2	63.0	55.2	39.2	45.7	42.2	42.6	51.3	46.6	39.7	56.5	46.7
ERGO(Chen et al., 2022)	49.7	72.6	59.0	43.2	48.8	45.8	46.3	50.1	48.1	58.4	60.5	59.4
CHEER(Chen et al., 2023)	56.9	69.6	62.6	45.2	52.1	48.4	49.7	53.3	51.4	56.4	69.5	62.3
SENDIR(Yuan et al., 2023)	65.8	66.7	66.2	33.0	90.0	48.3	37.8	82.8	51.9	65.2	57.7	61.2
PPAT(Liu et al., 2023c)	62.1	68.8	65.3	54.0	50.2	52.0	56.8	56.0	56.4	62.5	62.4	62.4
iLIF(Liu et al., 2024)	76.8	66.3	71.2	53.5	65.9	59.1	59.2	66.1	62.5	-	-	-
EHNEM(Xiang et al., 2025a)	63.2	70.4	66.6	62.3	59.9	61.0	62.6	63.1	62.8	-	-	-
LLaMA-2-7B	16.8	98.0	28.7	6.11	98.3	11.5	7.1	98.6	14.2	3.8	93.3	7.4
LLaMA-3.1-8B	20.6	94.5	33.8	8.6	93.8	15.7	10.5	94.0	19.0	4.7	90.6	8.9
Qwen-2.5-7B	18.4	96.3	30.9	7.2	95.9	13.4	8.9	96.1	16.4	5.9	57.6	10.7
Qwen-2.5-14B	27.1	82.8	40.8	12.2	80.6	21.2	14.9	81.3	25.1	5.0	88.9	9.5
Deepseek-v3.2	33.3	66.7	44.4	18.2	68.9	28.8	21.2	68.3	32.3	6.1	86.9	11.4
SVRECI	**79.0**	66.8	**71.9**	60.0	63.5	**61.9**	**65.5**	64.5	**65.0**	**66.5**	63.1	**64.5**

Table 1: Experimental Results on the ESC and CTB Datasets (%).

Methods	Cross			Intra & Cross		
	P	R	F1	P	R	F1
SVRECI	**60.0**	**63.5**	**61.9**	**65.5**	**64.5**	**65.0**
Suggest stage						
-SG	56.5	61.3	58.8	62.2	61.1	61.6
-SG$_{LLM}$	58.8	62.1	60.4	64.0	62.3	63.1
-SG$_{Hyp}$	57.6	61.5	59.5	63.1	61.8	62.4
Verify stage						
-TV	57.8	61.9	59.8	63.0	61.5	62.2
Revise stage						
-IR	55.6	60.8	58.1	61.2	60.5	60.8
-CL	58.4	59.1	58.7	63.4	60.8	62.1
-CL$_{event}$	59.1	60.0	59.5	64.2	61.6	62.9
-CL$_{pair}$	59.6	60.4	60.0	64.9	62.0	63.6

Table 2: Ablation results on ESC (%).

Methods	Prop.	Prun.	Iter.	C-F1	O-F1
SVRECI	✓	✓	✓	**61.9**	**65.0**
-prop.		✓	✓	59.2	62.5
-prun.	✓		✓	59.3	63.2
-iter.	✓	✓		58.7	61.5
-IR				58.1	60.8

Table 3: Fine-grained ablation of the IR module.

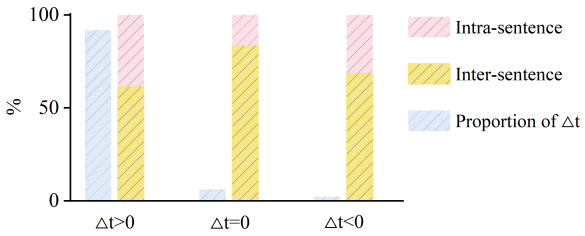

Figure 3: Narrative consistency of causal pairs on the ESC dataset.

reported in Table 2, respectively. Removing any individual component consistently degrades performance, while replacing iterative refinement with a single-step update results in the largest drop among partial variants. Removing the entire IR module causes the most significant degradation, confirming that these components provide complementary gains and that iterative updates are essential for progressively refining the causal graph.

4.5 Analysis of Causal Role Assignment

We analyze the impact of causal role assignment on directionality identification on ESC dataset, as shown in Table 4. SVRECI achieves the best performance in intra-sentence, cross-sentence, and overall settings, demonstrating that explicitly modeling the functional roles of events in causal chains and enforcing role consistency at the structural level mitigates common role confusion and direction reversal in cross-sentence scenarios. In contrast, LLM-based generative methods show markedly lower precision for directionality prediction, reflecting the lack of verifiable structural causal constraints. These results suggest that causal role assignment provides a crucial structural inductive bias for stable and reliable document-level directionality inference.

4.6 Narrative Consistency Analysis

We analyze the narrative consistency of SVRECI on the ESC dataset by computing the narrative

Methods	EventStoryLine								
	Intra			Cross			Intra and Cross		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BERT (Devlin et al., 2019)	62.4	32.6	42.8	34.4	30.7	32.4	40.7	31.3	35.4
RoBERTa(Liu et al., 2019)	59.7	38.0	46.4	31.3	34.2	32.7	37.3	35.5	36.4
LONG(Beltagy et al., 2020)	59.0	40.5	48.0	35.2	30.5	32.7	41.6	33.8	37.3
ERGO(Chen et al., 2022)	58.8	47.6	52.6	36.1	41.2	38.5	41.5	43.3	42.4
SENDIR(Yuan et al., 2023)	56.0	52.6	54.2	38.6	39.4	39.0	43.8	43.7	43.7
iLIF (Liu et al., 2024)	66.7	54.5	60.0	41.2	44.6	42.8	47.9	47.8	47.8
LLaMA-2-7B	8.4	41.3	14.0	3.2	44.3	6.0	4.0	43.3	7.2
LLaMA-3.1-8B	11.1	56.4	18.6	3.7	52.4	6.8	4.7	53.7	8.6
Qwen-2.5-7B	9.3	49.9	15.7	3.5	54.4	6.6	4.3	52.9	8.0
Qwen-2.5-14B	10.6	53.5	17.7	4.0	54.0	7.4	5.0	53.8	9.2
Deepseek-v3.2	16.4	75.2	26.9	8.7	69.4	15.4	10.3	71.3	18.0
SVRECI	**68.6**	56.0	**61.8**	**46.2**	44	**45.1**	**52.7**	47.8	**50.2**

Table 4: Performance comparison on the ESC in direction evaluation settings.

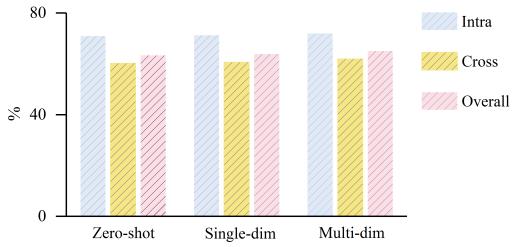

Figure 4: Performance under different prompt templates of increasing complexity.

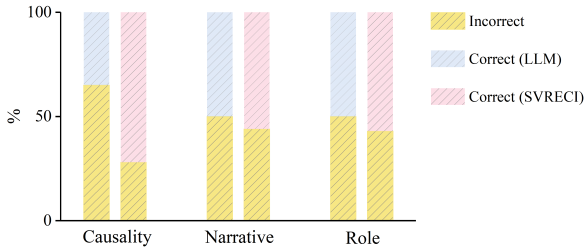

Figure 5: Human evaluation results across different assessment dimensions.

temporal difference $\Delta t = t_j - t_i$ for each predicted causal pair (e_i, e_j), as shown in Figure 3. We observe that the vast majority of causal relations predicted by SVRECI conform to the narrative propagation order, with cross-sentence causal relations accounting for a substantial proportion. This indicates SVRECI's effectiveness in uncovering cross-sentence causal propagation along the document-level narrative backbone. For the case where $\Delta t = 0$, the predictions are mainly concentrated on intra-sentence relations, which is attributable to the use of sentence-level order as a proxy in the absence of explicit temporal annotations. Overall, the causal structures learned by SVRECI exhibit strong sequential consistency and structural plausibility from a global narrative propagation perspective.

4.7 Prompt Sensitivity Analysis

We evaluate the sensitivity of LLM-generated heuristic suggestions to prompt design by comparing three templates of increasing complexity: zero-shot prompts, single-dimension heuristic prompts (see Appendix B.1), and multi-dimensional structured prompts (see Appendix B). The zero-shot prompt is illustrated in Figure 15, and the results (F1 scores) are shown in Figure 4, where Overall denotes the Intra-and-Cross setting.

As the prompt becomes more informative, performance consistently improves across all settings. More importantly, even with the simplest zero-shot prompt, the model still consistently outperforms baseline methods after the verification and refinement stages. This indicates that, although performance is influenced by prompt design, the overall framework effectively mitigates local prompt noise through global consistency modeling and iterative optimization.

4.8 Human Evaluation Analysis

As shown in Figure 5, we randomly sample 100 event pairs and evaluate LLM initial suggestions and SVRECI-refined results from three aspects: causal correctness, narrative consistency, and role plausibility (see Appendix J). The results indicate that, despite using multi-dimensional heuristic prompts, LLM initial suggestions contain errors

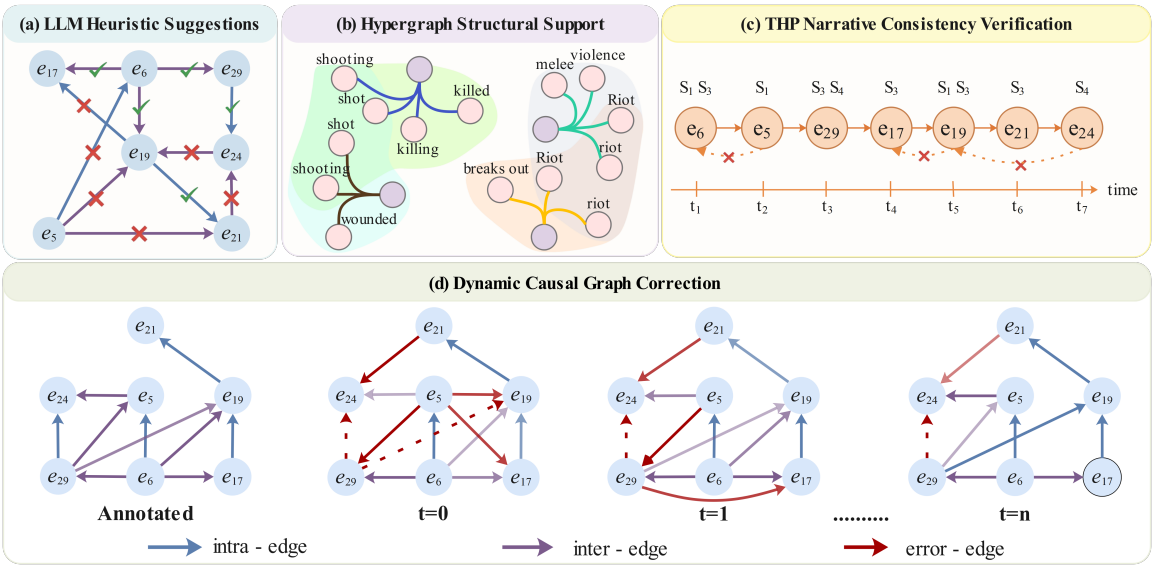

Figure 6: Case study. (a) LLM heuristic suggestions; (b) Hypergraph-based structural suggestions via event coreference and concept-level aggregation (different colors denote different hyperedges); (c) Narrative consistency verification via THP, where globally inconsistent edges are progressively suppressed; (d) Iterative causal graph refinement, where color intensity indicates causal strength, and red edges denote errors.

in all three dimensions, whereas SVRECI refinement substantially improves accuracy, validating the effectiveness of our model.

4.9 Case Study

Figure 1(a) presents a real document with 29 events across 4 sentences, which serves as the running example for document-level causal identification. Figure 6(a) shows that, even with carefully designed prompts, the initial suggestions from the LLM still contain some errors. Figure 6(b) presents hypergraph-based structural suggestions derived from event coreference and concept aggregation. Figure 6(c) shows the THP-based narrative consistency verification, which progressively suppresses noisy edges via propagation-aware reweighting. Figure 6(d) illustrates the iterative graph revision process, where erroneous edges are gradually corrected through consistency verification after each iteration, effectively improving the accuracy of the document-level causal structure.

5 Conclusion

In this paper, we propose a suggest-verify-revise approach for document-level event causality identification with narrative consistency. SVRECI unifies LLM-based heuristic suggestions and hypergraph structural proposals, verifies candidate causal edges via topological Hawkes processes for global narrative consistency, and constructs a dynamically evolving causal graph refined by structure-aware dual-level contrastive learning. Experiments on two widely used datasets demonstrate that our approach achieves strong performance on the ECI.

6 Limitations

In this work, the heuristic suggestions generated by the LLMs are sensitive to prompt design, and different prompt templates may lead to fluctuations in suggestion quality. Additionally, in datasets lacking explicit temporal annotations, we approximate event narrative time using sentence-level order, potentially introducing temporal noise and affecting the accuracy of narrative consistency verification. Future work will explore more robust and interpretable reasoning frameworks (Fan et al., 2025) under diverse real-world narrative settings, as well as more efficient and scalable multi-agent or collaborative architectures (Zhang et al., 2026, 2024).

Acknowledgements

We thank all the anonymous reviewers for their constructive comments and suggestions. This work is supported by the National Natural Science Foundation of China (62476161, 62176145), the Interdisciplinary Research Fund of Shanxi University, the Natural Language Processing Innovation Team (Sanjin Talents) Project of Shanxi Province.

References

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2021. Automated storytelling via causal, commonsense plot ordering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5859–5867.

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *International conference on intelligent text processing and computational linguistics*, pages 430–441. Springer.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. 2021. Thps: Topological hawkes processes for learning causal structure on event sequences. *IEEE Transactions on Neural Networks and Learning Systems*, 35:479–493.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

Liang Chao, Wei Xiang, and Bang Wang. 2024. In-context contrastive learning for event causality identification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 868–881.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. Cheer: Centrality-aware high-order event reasoning network for document-level event causality identification.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ling Ding, Jianting Chen, Peng Du, and Yang Xiang. 2024. Event causality identification via graph contrast-based knowledge augmented networks. *Information Sciences*, 656:119905.

Yue Fan, Hu Zhang, Ru Li, Yujie Wang, Guangjun Zhang, Hongye Tan, and Jiye Liang. 2025. Weakly-supervised explainable question answering via question aware contrastive learning and adaptive gate mechanism. *Information Sciences*, 697:121763.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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)*, pages 1808–1817.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Shunhang Li, Gang Zhou, Jing Chen, Yepeng Sun, Ningbo Huang, and Sisi Peng. 2025. Event-level supervised contrastive learning with back-translation augmentation for event causality identification. *Neurocomputing*, 621:129232.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Cheng Liu, Wei Xiang, and Bang Wang. 2024. 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)*, pages 3815–3827, Bangkok, Thailand. Association for Computational Linguistics.

Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023a. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-based systems*, 259:110064.

Yang Liu, Guanbin Li, and Liang Lin. 2023b. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhenyu Liu, Baotian Hu, Zhenran Xu, and M. Zhang. 2023c. Ppat: Progressive graph pairwise attention

network for event causality identification. In *International Joint Conference on Artificial Intelligence*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19.

Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. 2017. Multi-column convolutional neural networks with causality-attention for why-question answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 415424, New York, NY, USA. Association for Computing Machinery.

Minh Tran Phu and Thien Huu Nguyen. 2021. 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*, pages 3480–3490.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 909918, New York, NY, USA. Association for Computing Machinery.

Wen-hui Sah. 2013. The development of coherence in narratives: Causal relations. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 173–180.

Ya Su, Hu Zhang, Yue Fan, Guangjun Zhang, Yujie Wang, Ru Li, and Hongye Tan. 2025a. Dynamic energy-based contrastive learning with multi-stage knowledge verification for event causality identification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12260–12278.

Ya Su, Hu Zhang, Guangjun Zhang, Yujie Wang, Yue Fan, Ru Li, and Yuanlong Wang. 2025b. Enhancing event causality identification with llm knowledge and concept-level event relations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7403–7414.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Tom Trabasso and 1 others. 1982. Causal cohesion and story coherence.

Sifan Wu, Ruihui Zhao, Yefeng Zheng, Jian Pei, and Bang Liu. 2023. Identify event causality with knowledge and analogy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13745–13753.

Wei Xiang, Cheng Liu, and Bang Wang. 2025a. Modeling correlated causal-effect structure with a hypergraph for document-level event causality identification. *Computer Speech & Language*, 90:101752.

Wei Xiang, Chuanhong Zhan, Qing Zhang, and Bang Wang. 2025b. Daprompt: Deterministic assumption prompt learning for event causality identification. *Neural Computing and Applications*, 37(26):21743–21759.

Hao Xu, Yunxiao Zhao, Jiayang Zhang, Zhiqiang Wang, and Ru Li. 2025. LOG: A local-to-global optimization approach for retrieval-based explainable multi-hop question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9085–9095, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. 2016. Learning granger causality for hawkes processes. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1717–1726, New York, New York, USA. PMLR.

Changsen Yuan, He-Yan Huang, Yixin Cao, and Yonggang Wen. 2023. Discriminative reasoning with sparse event representation for document-level event-event relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16222–16234.

Guangjun Zhang, Hu Zhang, Yazhou Han, Yue Fan, Yuhang Shao, Hongye Tan, and Ru Li. 2026. Learning to generate and extract: A multi-agent collaboration framework for zero-shot document-level event arguments extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34665–34673.

Guangjun Zhang, Hu Zhang, YuJie Wang, Ru Li, Hongye Tan, and Jiye Liang. 2024. Hyperspherical multi-prototype with optimal transport for event argument extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9271–9284, Bangkok, Thailand. Association for Computational Linguistics.

Yunxiao Zhao, Zhiqiang Wang, Xingtong Yu, Xiaoli Li, Jiye Liang, and Ru Li. 2026. Learnable game-theoretic policy optimization for data-centric self-explanation rationalization. *IEEE Transactions on Knowledge and Data Engineering*, 38(2):1159–1173.

A Event Conceptual Aggregation and Multi-Cause Multi-Effect Role Constraints

Aggregation of Similar Event Mentions into Event Concepts: As shown in Figure 1(a), multiple semantically similar but lexically different event mentions exist in the text (e.g., anger/Anger, riot/Riot, followed/following, killed/killing), which collectively refer to the same conceptual event. This concept aggregation structurally forms multi-perspective anchors on the narrative backbone, helping maintain the integrity of document-level causal chains. As shown in Figure 1(b), the causal chains formed by mentions such as anger/Anger and riot/Riot jointly support the narrative causal backbone.

Multi-Cause Multi-Effect Shared Structure: In real-world documents, causal relations are not determined by independent event pairs; events generally participate in multi-cause and multi-effect shared structures. A single event may serve as a common cause for multiple subsequent events and may also be triggered by multiple prior causes. For instance, as shown in Figure 1(c), event e_{19} can trigger e_{20}, e_{21}, and e_{22}, while e_6, e_{29}, e_{17}, and e_{16} jointly trigger e_{19}. This structure indicates that causal judgment is not an independent pairwise decision but is systematically influenced by higher-order event structures..

Event Role Diversity and Structural Functionality: The same event may assume different structural roles in different causal chains. For example, in the causal chain $e_{17} \rightarrow e_{19} \rightarrow e_{20}$, e_{19} functions as an intermediate node, whereas in $e_1 \rightarrow e_6 \rightarrow e_{19}$, e_{19} serves as a result node. Such role diversity implies that the causal support of an event pair must be modeled with respect to its relative position and functional role within the overall document structure, rather than relying solely on local semantics or distance information.

B Detailed LLM-based Heuristic Rules

Given an event pair (e_i, e_j) and its context, we generate multi-perspective reasoning texts based on chain-of-thought (CoT) using DeepSeek LLM (Liu et al., 2025). Heuristic evidence functions are then extracted from these reasoning texts, which are mapped to confidence scores in the range $[0, 1]$ to reflect the potential causal strength from different perspectives:

$$h(e_i, e_j) = \left[h_{ij}^{\text{sem}}, h_{ij}^{\text{fun}}, h_{ij}^{\text{dep}}, h_{ij}^{\text{temp}}\right], \quad h_{ij}^{k} \in [0, 1] \tag{16}$$

B.1 Semantic Heuristic

Semantic heuristics focus on whether two events exhibit an explicit "Cause $\rightarrow$ Effect" conceptual mapping at the semantic level, rather than mere topical relevance or background co-occurrence. As shown in Equation 17, this heuristic is modeled by analyzing the semantic cues within the LLM-generated reasoning text and the original sentences.

$$h_{ij}^{sem} = \begin{cases} \alpha_1, & \text{if explicit causality} \\ \alpha_2, & \text{if consequence-oriented} \\ & \text{dependency expressed} \\ 0, & \text{if only background correlation} \end{cases} \tag{17}$$

In Equation 17, α_1 and α_2 represent the weights for the following rule-based designs:

- **Explicit Causal Semantic Triggering:** If explicit causal indicators (e.g., *cause*, *lead to*, *result in*) are detected in the sentences or LLM reasoning, the event pair is assigned a high confidence score α_1.
- **Consequence Dependency Expression:** If event e_j is semantically described as the consequence, purpose, or response to e_i, the causal confidence is reinforced with α_2.
- **Correlation Trap Suppression:** If two events only share external contexts (e.g., the same person or policy) without an explicit dependency, the heuristic value is set to 0 to suppress potential false positives.

Detailed prompt designs are further elaborated in Figure 7.

B.2 Causal Role Heuristic

The causal role heuristic models the potential causal-explanatory relationships between events by analyzing syntactic functional structures. Unlike methods relying on formal dependency trees, this heuristic captures the underlying logic of "what triggers what" and the specific roles events play within causal predicates. As shown in Equation 18:

$$h_{ij}^{\text{cau}} = \mathbf{I}\left(e_i \xrightarrow{\text{con / sub / ver}} e_j\right) \tag{18}$$

where $\mathbf{I}(\cdot)$ is an indicator function governed by the following rules:

- **Causal Role Connection (*con*):** Confidence is increased if events are linked via causal connectives (e.g., *because*, *due to*).
- **Subordinate Explanatory Relation (*sub*):** Causal potential is identified if a subordinate clause functionally explains the main clause event.
- **Causal Role Constraint (*ver*):** Causal strength is reinforced if e_i serves as the agent of a causal verb and e_j acts as its resultative patient or complement.

Detailed prompt designs and examples are provided in Figure 8 and Figure 9.

B.3 Dependency Parsing Heuristic

Dependency analysis heuristic focuses on the structural causal patterns of event triggers in the dependency tree. We define potential causal paths between events on the dependency tree $\mathcal{T}$, and particularly consider the following dependency pattern:

$$h_{ij}^{\text{dep}} = I(\text{causal dependency path in } \mathcal{T}) \quad (19)$$

where the key dependency patterns include:

- **Subject-Verb-Object Causal Pattern**: `nsubj` + causal verb (*cause*, *trigger*, *lead*) + `dobj`
- **Adverbial Clause Causal Pattern**: `advcl` indicating cause or effect
- **Prepositional Causal Phrase**: `prep` connecting causal phrases such as *due to*, *because of*, etc.

If an event pair does not have a direct or indirect causal dependency path in the dependency tree, this heuristic does not positively adjust the causal strength. Corresponding prompt designs and examples are provided in Figure 10,Figure 11 and Figure 12.

B.4 Temporal Order Heuristic

In event causality identification tasks, temporal order is a necessary but not sufficient condition for causality. We determine the chronological relationship between events based on explicit or implicit temporal cues:

$$h_{ij}^{tmp} = \begin{cases} 1, & e_i <_t e_j \\ 0, & e_j <_t e_i \end{cases} \quad (20)$$

where the chronological logic implies:

- If event e_i precedes event e_j in time, the heuristic is positive.
- If the temporal order is reversed, the causal strength is suppressed.
- If only temporal order exists without other semantic or structural evidence, this heuristic does not independently increase the overall causal strength.

Corresponding prompt designs and examples are provided in Figure 13 and Figure 14.

B.5 Aggregation of LLM Heuristic Causal Strength

The above heuristic evidence is ultimately aggregated to form the LLM heuristic causal strength, which serves as a soft prior input for subsequent graph-structured modeling and iterative refinement:

$$s_{ij}^{\text{LLM}} = \frac{1}{K} \sum_{k=1}^{K} h_{ij}^{k} \quad (21)$$

C Multi-dimensional Features of the Structural Hypergraph

In real narratives, causal relations are not determined as independent event pairs, but are embedded in shared narrative structures. An event may act as a common source or a common effect, participating in multiple potential causal chains. Therefore, for any event pair (e_i, e_j), a multidimensional structural support vector is extracted from the document-level structural hypergraph:

$$\mathbf{f}_{ij}^{\text{struct}} = [f_{ij}^{(1)}, f_{ij}^{(2)}, f_{ij}^{(3)}] \in \mathbb{R}^3 \quad (22)$$

This vector characterizes the potential support of the event pair in the structural hypergraph. It is not used for direct causal judgment, but serves as a structural soft signal input to the Hawkes process. The meanings of each dimension are as follows:

Coreference Shared Strength $f_{ij}^{(1)}$: Measures the frequency of co-occurrence of the event pair in coreference shared structures, reflecting the degree

of conceptual binding of the events within the structure. Higher frequency indicates stronger potential association of the event pair in structural sharing.

$$f_{ij}^{(1)} = \frac{1}{|\mathcal{E}_S|} \sum_{h \in \mathcal{E}_S} \mathbb{I}(e_i \in h \wedge e_j \in h) \tag{23}$$

Structural Positional Discrepancy $f_{ij}^{(2)}$: Characterizes the functional positional difference of events in the document-level narrative, rather than simple textual distance. This metric acts as a weak structural constraint to suppress pseudo-causal candidates that exhibit clearly unreasonable narrative positions, while not negating reasonable cross-sentence or cross-paragraph true causal relations. We map each event e to a low-dimensional positional encoding vector:

$$\phi(e) = [\phi_{\text{sent}}(e), \phi_{\text{para-proxy}}(e)] \tag{24}$$

where $\phi_{\text{sent}}(e)$ is the normalized relative position of the event within the sentence, and $\phi_{\text{para-proxy}}(e)$ is the normalized code of its functional paragraph. The paragraph proxy is determined according to the sentence t of the event and total number of sentences T in the document, dividing into Intro, Development (Dev), and Outcome segments:

$$\pi(s_t) = \begin{cases} \text{Intro}, & t \leq 0.2T \\ \text{Dev}, & 0.2T < t \leq 0.7T \\ \text{Outcome}, & t > 0.7T \end{cases} \tag{25}$$

Let the set of positional dimensions be $\mathcal{R} = \{\text{sent}, \text{para-proxy}\}$. Then the structural discrepancy of an event pair (e_i, e_j) is defined as:

$$f_{ij}^{(2)} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} |\phi_r(e_i) - \phi_r(e_j)| \tag{26}$$

Global Structural Centrality $f_{ij}^{(3)}$: Measures the exposure and potential anchoring role of events in the structural hypergraph. Events participating in multiple structural sharing relations (e.g., coreference clusters or shared structural units) are more likely to be located at intersections of potential causal paths, thus acting as structural anchors in subsequent causal reasoning. This metric is computed via hypergraph degree and extended to the event pair level. It does not directly reflect causal relations but provides a global structural prior to assist causal consistency verification.

$$f_{ij}^{(3)} = \frac{\text{Centrality}(e_i) + \text{Centrality}(e_j)}{2}, \tag{27}$$
$$\text{Centrality}(e) = |\{h \in \mathcal{E}_S \mid e \in h\}|$$

D Motivation for Narrative Consistency and Topological Hawkes Modeling

Generativity of Causality and Narrative Role Differentiation: Existing ECI methods (Chen et al., 2022; Liu et al., 2024; Xiang et al., 2025a) mostly employ discriminative models to classify event pairs independently, relying primarily on local semantic similarity. However, in real narrative texts, causal relationships are not symmetric semantic similarity relations but manifest as directional explanatory structures. The rising/falling action annotations in ESC v0.9 imply a clear narrative explanatory direction, indicating that certain events are more likely to assume the role of an explanatory source (cause-like), while others fit the role of a result (effect-like). Therefore, modeling event causality identification as a discriminative task of independent event pairs easily leads to misidentifying events that are parallel, co-occurring, or sharing antecedents as causally linked, especially in multi-cause-to-single-effect and cross-sentence narrative scenarios. This suggests that event causality identification requires not only local semantic judgment but also document-level structural consistency constraints.

Narrative Time as Weak Supervision Signals: Although ESC v0.9 does not provide continuous timestamps, its TLINK annotations and document narrative order provide reliable weak supervision signals for the relative temporal relations of events. We abstract these signals into discrete narrative time anchors to characterize the relative positions of events within the document structure. Considering that narrative order does not always strictly correspond to actual causal order, we treat it as a soft prior rather than a hard constraint: candidate causal edges that violate the narrative order are not directly excluded but are progressively weakened in a continuous space, allowing subsequent modules to perform corrections.

Based on these observations, in Section3.2, we introduce the Topological Hawkes Process (THP) as a narrative consistency verification module to characterize the overall rationality of candidate causal relationships under the constraints of document topology and narrative order. THP does not fit actual occurrence times; instead, it performs continuous and differentiable global consistency modeling of candidate causal structures under discrete narrative time and document structure constraints, thereby providing a structural adjustment

mechanism for document-level causal graphs.

E Structure-Aware Heterogeneous Graph Propagation and Self-Optimization Mechanism

Structure-Aware Heterogeneous Graph Attention Propagation: In each optimization iteration, if the current causal graph contains valid candidate edges, a Heterogeneous Graph Convolutional Network (HGCN) based on an attention mechanism is employed to update event representations; otherwise, the model degrades to a linear update based solely on node semantics to ensure stability in weak or empty graph scenarios.

In the l-th layer, the representation of event node e_i is $\mathbf{h}_i^{(l)}$. For a candidate causal edge $e_i \to e_j$, we first calculate the attention score determined solely by semantics:

$$e_{i\to j}^{(l)} = \text{LeakyReLU}\Big(\mathbf{a}_{\tau_{(i,j)}}^{\top}[\mathbf{W}_{\tau_{(i,j)}}\mathbf{h}_i^{(l)} \| \mathbf{W}_{\tau_{(i,j)}}\mathbf{h}_j^{(l)}]\Big) \tag{28}$$

where $\tau_{(i,j)} \in \{\text{intra, inter}\}$ denotes the edge type; $\mathbf{W}_\tau$ and $\mathbf{a}_\tau$ are type-specific learnable parameters. To incorporate structural confidence information while avoiding the disruption of the relative ranking of semantic attention, we do not inject edge weights directly as additive or multiplicative terms into the attention score. Instead, we map them into a confidence gating factor:

$$\gamma_{i\to j}^{(l)} = \sigma(\mathbf{W}_g \cdot w_{i\to j}^{(l)}) \tag{29}$$

And control the intensity of information flow via multiplicative modulation:

$$\alpha_{i\to j}^{(l)} = \text{softmax}_i\left(e_{i\to j}^{(l)} \cdot \gamma_{i\to j}^{(l)}\right) \tag{30}$$

This design possesses a ranking-preserving property, where semantic attention determines "whom to focus on," while the structural prior only regulates the "strength of information transmission".

Multi-head Propagation and Type-level Fusion: Under the multi-head attention mechanism, the updated representation of an event node is:

$$h_j^{(l+1)} = \sum_{m=1}^{M} \sum_{(i\to j)\in\mathcal{E}_c^{(t)}} \alpha_{i\to j,m}^{(l)} \cdot W_{m,\tau} h_i^{(l)} \tag{31}$$

When a node simultaneously receives information from intra-sentence and inter-sentence causal edges, we further introduce type-level fusion:

$$\mathbf{h}_j^{(l+1)} = \eta \cdot \mathbf{h}_{j,\text{intra}}^{(l+1)} + (1-\eta)\cdot \mathbf{h}_{j,\text{inter}}^{(l+1)}, \quad \eta \in (0,1) \tag{32}$$

where η is a learnable parameter used to highlight the relative stability of intra-sentence causal relations in document understanding, while preserving the capability for cross-sentence causal reasoning.

To prevent graph structural features from overwhelming the node's own semantics and to enhance robustness in weak-structure scenarios, we adopt gated residual fusion:

$$g_j = \sigma\left(\mathbf{W}_g\left[\mathbf{h}_{j,\text{ self}} \| \mathbf{h}_{j,\text{ graph}}\right]\right) \tag{33}$$

$$h_j^{(l+1)} = g_j \odot h_{j,\text{graph}} + (1-g_j)\odot h_{j,\text{self}} \tag{34}$$

The gating vector g_j is learned from the node itself, with LayerNorm subsequently applied to enhance training stability.

F Dual-Level Structure-Aware Contrastive Learning Mechanism

To strengthen the consistency of causal structures and boundary discrimination, we propose a dual-level iteration-aware contrastive learning strategy. In each iteration, positive and negative samples of events and event pairs are dynamically sampled based on the current causal graph to avoid mismatch between structure and representation caused by static sampling. Unlike existing methods (Gao et al., 2023), we introduce structural contrastive learning for the first time, which does not rely on semantics to construct positive and negative examples, making it suitable for scenarios where positive examples are scarce.

Event-Level Structural Contrastive Learning: Event-level contrastive learning is used to regularize the structural role consistency of individual events within the causal system. Given an event e_i, its structural representation $\mathbf{s}_i$ is composed of its 1-hop causal subgraph:

$$\mathbf{s}_i = \left[\mathbf{h}_i \| \frac{1}{|N(i)|}\sum_{j\in N(i)} \mathbf{h}_j \| \frac{1}{|N(i)|}\sum_{(i,j)} w_{ij}\right] \tag{35}$$

where $\mathbf{h}_i$ is the representation of the event itself, $N(i)$ is the set of neighboring events, and w_{ij} is the edge confidence in the current iteration. If $N(i) = \emptyset$, the neighbor and edge weight parts are set to zero. The contrastive vector is obtained after a learnable mapping and normalization:

$$\mathbf{z}_i = \text{Normalize}(\text{Proj}(\mathbf{s}_i)) \tag{36}$$

The causal roles of events are implicitly induced by the document-level causal graph (e.g., cause-like

/ effect-like). For each event acting as an anchor, positive samples are selected from events with consistent roles, integrating semantic and structural similarity:

$$\text{sim}_{\text{pos}}(i,j) = \cos(\mathbf{h}_i, \mathbf{h}_j) \cdot \frac{1}{1+|\deg(i)-\deg(j)|} \tag{37}$$

Negative samples include: (1) Hard negative samples, which are the most structurally similar events selected from those with opposite roles; (2) Randomly sampled easy negative samples. The event-level contrastive loss employs a weighted InfoNCE:

$$\begin{aligned}\mathcal{L}_{\text{event}} = -\log \frac{\exp(\cos(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\exp(\cos(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)} \\ +2\exp(\cos(\mathbf{z}_i, \mathbf{z}_i^{\text{hard}-})/\tau) \\ +\exp(\cos(\mathbf{z}_i, \mathbf{z}_i^{\text{rand}-})/\tau)\end{aligned} \tag{38}$$

This objective imposes structural role consistency regularization on event representations without directly predicting causal relations.

Event-Pair-Level Structural Contrastive Learning: To explicitly align structural representations with causal discrimination boundaries, we introduce event-pair-level structural contrastive learning, modeling causal relationships as compact relational units in a joint embedding space. For an event pair (e_i, e_j), its pair representation is defined as:

$$\mathbf{h}_{ij} = [\|\mathbf{z}_i\| \| \|\mathbf{z}_j\| \| \|\mathbf{z}_i - \mathbf{z}_j\| \| \|\mathbf{z}_i \odot \mathbf{z}_j\|] \tag{39}$$

where $\mathbf{z}_i, \mathbf{z}_j$ are structured event embeddings refined by the graph. Let $\mathcal{P}^+$ and $\mathcal{P}^-$ denote the sets of causal and non-causal event pairs, respectively. For any non-causal pair, its hardness is defined as:

$$\text{hardness}(i,j) = \max_{(k,l)\in\mathcal{P}^+} \cos(\mathbf{z}_{ij}, \mathbf{z}_{kl}) \cdot (1+w_{ij}) \tag{40}$$

The top-K pairs are selected as hard negatives $\mathcal{P}^-_{\text{hard}}$. The selection of positive and negative samples is as follows:

$$\mathbf{z}_{ij}^+ = \arg\max_{(k,l)\in\mathcal{P}^+\setminus(i,j)} \cos(\mathbf{z}_{ij}, \mathbf{z}_{kl}) \tag{41}$$

$$\mathbf{z}_{ij}^- = \arg\max_{(k,l)\in\mathcal{P}^-_{\text{hard}}} \cos(\mathbf{z}_{ij}, \mathbf{z}_{kl}) \tag{42}$$

A margin-based ranking loss is adopted:

$$\begin{aligned}\mathcal{L}_{\text{pair}} = \sum_{(i,j)\in\mathcal{P}^+} \max(0, \gamma + \cos(\mathbf{z}_{ij}, \mathbf{z}_{ij}^-) \\ -\cos(\mathbf{z}_{ij}, \mathbf{z}_{ij}^+))\end{aligned} \tag{43}$$

This objective explicitly pushes away non-causal structures that are most similar to causal ones, thereby strengthening the discrimination boundary and suppressing representation collapse. The structural contrastive loss is enabled only after the causal graph has initially converged and is introduced gradually through weight scheduling to reduce the interference of early noisy structures on the main task.

G Datasets

This study utilizes two public document-level event causality identification (ECI) datasets: EventStoryLine v0.9 (ESC) and Causal-TimeBank (CTB). We conduct a statistical analysis of the data features to reveal the practical challenges faced when constructing contrastive learning samples.

ESC: This dataset contains 22 topics, 258 documents, and a total of 5,334 event mentions. Among them, there are 7,805 intra-sentence event pairs, including 1,770 causal pairs (22.67%), and 46,521 inter-sentence event pairs, including 3,855 causal pairs (8%). The total number of causal links in the dataset is 5,625, of which only 117 are explicit causal links, with the rest being implicit. Furthermore, 34% of event pairs with causal relations have coreference relations. We primarily identify the `PRECONDITION` tag as the `CAUSE` relation, the `FALLING_ACTION` tag as the `EFFECT` relation, and utilize `TLINK` annotations as timestamps. This dataset supports both intra-sentence and inter-sentence event causality identification.

CTB: This dataset consists of 183 documents and 6,811 event mentions. There are 9,721 intra-sentence event pairs, including 298 causal pairs (3.1%), and approximately 7,608 inter-sentence event pairs, but these contain only 18 causal relations (0.24%), with all spans not exceeding three sentences. Notably, the number of inter-sentence causal pairs is extremely limited, with only 20 found among 252,084 total inter-sentence pairs. Overall, the ratio of positive to negative samples is approximately 1:3 for intra-sentence pairs and 1:10 for inter-sentence pairs. Due to the extreme sparsity of inter-sentence causal samples, we follow previous studies (Gao et al., 2023; Chao et al., 2024; Su et al., 2025b) and restrict our evaluation of sentence-level event causality identification to the Causal-TimeBank dataset.

Overall, high-quality causal positive examples are scarce in both datasets, which explains why

existing contrastive learning methods (Loshchilov and Hutter, 2017) are susceptible to noise interference and exhibit unstable performance during sample construction. To address this, this study adopts a structural contrastive learning strategy to replace traditional methods based on random or heuristically constructed sample pairs, thereby enhancing the model's robustness in sparse and noisy environments and improving its structural consistency.

H Hyperparameter Settings

The model utilizes BERT-base-uncased as the encoder with a hidden dimension of 768. Event representations are obtained by performing mean pooling on the vectors of all tokens within their mention spans. To ensure reproducibility, the random seed is fixed at 209.

Dynamic Causal Graph Encoder: We construct a dynamic heterogeneous causal graph comprising both intra-sentence and inter-sentence edges. Based on Graph Attention Networks, we design a causal-aware heterogeneous graph attention encoder that executes relation-specific multi-head attention calculations through dynamic edge weight adjustment and gated residual fusion mechanisms, facilitating iterative optimization of the document-level causal structure. The fusion weights for intra-sentence and inter-sentence relations are set to 0.7 and 0.3, respectively. The output dimension of the graph encoder is 768, with 4 attention heads. The graph structure optimization is controlled via a self-iteration mechanism, with a minimum of 2 and a maximum of 9 iterations. The structural convergence threshold is set to 2.

Classification and Training Objectives: We consider two tasks: causal existence identification (CAUSE/NONE) and causal role classification (CAUSE/EFFECT/NONE). In the role classification task, the weights for the CAUSE and EFFECT categories are set to $\alpha = 0.75$, while the weight for the NONE category is $1 - \alpha$.

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{event}} + \lambda_2 \mathcal{L}_{\text{pair}} \tag{44}$$

where $\mathcal{L}_{\text{cls}}$ is implemented using Focal Loss with a focusing factor $\gamma = 2.0$ to alleviate class imbalance. To reinforce structural consistency, we introduce a dual-level structure-aware contrastive learning mechanism. For event-level contrastive loss, the temperature parameter is $\tau = 0.07$, the weight is $\lambda_1 = 0.1$, and it is introduced via an epoch-level warmup of 3. The event-pair-level contrastive loss is activated starting from epoch 10, with its weight λ_2 linearly increasing from 0.05. Edge weights across iterations are updated via EMA with a smoothing coefficient of 0.6. To reduce noise, structural contrastive loss is only enabled after the causal graph reaches initial convergence.

Optimization Settings: We employ AdamW as the optimizer with $\epsilon = 1e-8$, a gradient clipping threshold of 1.0, and a learning rate warmup ratio of 0.1.

I Baselines

Table 1 presents the experimental results of our model and various baselines on the EventStoryLine and Causal-TimeBank datasets. The baseline models are described as follows: RichGCN (Phu and Nguyen, 2021): Constructs a document-level graph integrating syntactic, semantic, and document structures, performing causal reasoning via GCN. It represents early graph-based methods for document-level causality identification.

ERGO (Chen et al., 2022): Proposes an Event Relational Graph Transformer that models multi-hop event dependencies through relation-aware self-attention.

CHEER (Chen et al., 2023): Introduces a centrality-aware higher-order reasoning mechanism, emphasizing structurally critical events and explicitly modeling higher-order causal paths.

SENDIR (Yuan et al., 2023): Performs discriminative reasoning based on sparse event representations.

PPAT (Liu et al., 2023c): Employs a progressive graph pair attention mechanism to iteratively optimize causal dependencies between event pairs.

iLIF (Liu et al., 2024): Proposes an "identifying while learning" paradigm, jointly optimizing causal graph derivation and relation classification.

EHNEM (Xiang et al., 2025a): Models correlated causal-effect structures through hypergraphs to achieve higher-order and multi-relational dependency modeling.

GCKAN (Ding et al., 2024): Introduces external knowledge graphs for knowledge enhancement, supplementing event semantics and potential reasoning knowledge.

ESupCL (Li et al., 2025): Utilizes back-translation augmentation to construct contrastive samples, increasing the representation gap between causal and

non-causal event pairs.

Furthermore, following the zero-shot setting in existing work (Gao et al., 2023; Liu et al., 2024), we introduce LLaMA-3.1 (7B and 8B) (Touvron et al., 2023), Qwen-2.5 (7B and 14B) (Hui et al., 2024), and DeepSeek (Liu et al., 2025) as LLM baselines. These models directly judge causal relations between event pairs through prompting without task-specific fine-tuning. Experiments demonstrate that LLM baselines typically exhibit high recall but low precision, with limited control over causal boundaries in zero-shot conditions. This further underscores the necessity and value of the "verification-correction" structured modeling framework proposed in this paper.

J Manual Evaluation Guidelines

To assess the quality of model-predicted event causal relations, we design a manual evaluation protocol that scores each event pair along three dimensions. Each dimension is assigned a value of 0 (incorrect) or 1 (correct) as follows:

Causality Correctness: Evaluates whether the predicted causal relation aligns with the textual semantics and factual logic. A score of 1 indicates the causal relation is correct, while 0 indicates an incorrect or unreasonable relation.

Narrative Consistency: Assesses whether the causal relation follows the document's narrative order and the propagation backbone of events. A score of 1 indicates the event order is consistent with the narrative, while 0 indicates conflicts or incoherence in the sequence.

Role Reasonability: Checks whether the event's role in the causal chain is appropriate. A score of 1 indicates the event's role aligns with its function in the narrative, while 0 indicates an unreasonable role, such as a result event being incorrectly assigned as a cause.

The final analysis counts the number of correct judgments in each dimension to evaluate model performance.

LLM Semantic Heuristic Prompt Design

Semantic Causal Heuristic Instruction

Role and Objective
You are a Causal Semantic Analyzer. Your task is to evaluate the semantic causal strength between two annotated events strictly based on the provided quantification rules. Your output will serve as heuristic evidence for downstream structured causal modeling; therefore, it must remain conservative, interpretable, and quantifiable.

Input Format
Sentence 1: "{sentence1}" Sentence 2: "{sentence2}"
Event 1: <t1>{event1}</t1> Event 2: <t3>{event2}</t3>
Note: Verify if the order <t1> $\rightarrow$ <t3> supports semantic causality.

Core Evaluation Rules (Three-Stage Quantitative Scoring)

Stage 1: Explicit Causal Semantic Trigger Identification (Wt 0.4)

Scan the text path between <t1> and <t3> for explicit causal indicators.

Score Criteria:

① 0.8–1.0 (Strong) "cause", "lead to", "result in", "trigger", "induce".
② 0.5–0.7 (Medium) "therefore", "so", "hence", "thus", "consequently".
③ 0.3–0.5 (Weak) "because", "due to", "for", "in order to".
④ 0.1 (None) No indicators found.
⑤ 0.0 (Negation) e.g., "did not cause".

Stage 2: Consequence Dependency Analysis (Wt 0.4)

Determine if <t3> is semantically the result, consequence, or target state of <t1>.

Score Criteria:

① 0.8–1.0 (Explicit) Direct outcome or inevitable consequence.
② 0.6–0.8 (Implicit) Logically holds via common sense/context, though not explicitly stated.
③ 0.3–0.5 (Related) Associated events but lacks strict result dependency.
④ 0.1–0.2 (Irrelevant) No logical connection.

Stage 3: Spurious Correlation Detection (Wt 0.2)

Detect if the link is a spurious correlation (events share background but lack causality).

Score Criteria:

① 0.0–0.2 (Confirmed Spurious) Only share Person, Time, Location, or Policy background.
② 0.3–0.5 (Potential Trap) Ambiguous connection.
③ 0.8–1.0 (No Trap) Clear causal mechanism exists; no spurious factors.

Calculation: $S_{\text{final}} = 0.4S_1 + 0.4S_2 + 0.2S_3$

Figure 7: The prompt design for semantic causal heuristic analysis.

LLM Causal Role Heuristic Prompt Design (Part 1)

Causal Role Structure Instruction

Role and Objective
You are a Causal Role Analyzer. Your task is to quantitatively evaluate whether there is a potential causal role dependency between two annotated events based on syntactic functional structures and event role cues.

Input Format
Sentence 1: "{sentence1}" Sentence 2: "{sentence2}"
Event 1: <t1>{event1}</t1> Event 2: <t3>{event2}</t3>
Note: The order <t1> → <t3> is a candidate order and must be supported by the syntactic structure itself.

Causal Role Heuristic Analysis Framework

Please analyze from the following three complementary causal role dimensions, and provide a sub-score in the [0,1] interval for each.

Dimension 1: Functional Connective Cues Detection (Wt 0.4)

Determine if there are explicit causal connectives or functional words directly or indirectly connecting <t1> and <t3>.

Scoring Reference

1. **Strong Causal Connectives (High Wt)**
   ① Cause/Causative (cause, lead to, result in, trigger, induce);
   ② Resultative (therefore, so, hence, thus);
   ③ Causal Prepositions (because of, due to).
2. **Medium Causal Connectives (Medium Wt)**
   ① Conditional (if...then, as long as, once);
   ② Purpose (in order to, so as to, for the purpose of, lest).
3. **Weak Causal Connectives (Low Wt)**
   ① Temporal Sequence (after, then, subsequently);
   ② Contrast (however, yet, but - only when implying unexpected results).

Quantitative Scoring Criteria

① 1.0: Strong causal connectives directly connecting the two events.
② 0.7: Strong causal connectives connecting indirectly.
③ 0.5: Medium causal connectives directly connecting.
④ 0.3: Weak causal connectives directly connecting.
⑤ 0.1: No relevant connectives found.
⑥ 0.0: Presence of negative causal expressions (e.g., "did not cause").

- -

Dimension 2: Causal Clausal Configuration Identification (Wt 0.3)

Identify if the two events are situated in a typical causal clause-main clause structure. Focus on the following configurations:

1. **Causal Adverbial Clause**
   ① Structure: Because/Since + Clause → Main Clause.
   ② example: "Because <t1> it rained</t1>, <t3> the match was canceled</t3>"
2. **Result Adverbial Clause**
   ① Structure: Main Clause → so that/such that + Clause.
   ② example: "The rain was so heavy that <t3> the road was flooded</t3>"

(Continued on next page...)

Figure 8: The prompt design for causal role heuristic analysis.

LLM Causal Role Heuristic Prompt Design (Part 2)

Causal Role Structure Instruction (Cont.)

Dimension 2 (Continued)

3. **Purpose Adverbial Clause**
   ① Structure: In order to/To + Clause → Main Clause.
   ② example: "To <t1> reduce risk</t1>, the company <t3> adjusted strategy</t3>"
4. **Conditional Adverbial Clause**
   ① Structure: If/Once + Clause → Main Clause.
   ② example: "If <t1> temperature exceeds threshold</t1>, system will <t3> shut down</t3>"

Quantitative Scoring Criteria:

① 1.0: Complete causal clause structure; events located in cause/result positions.
② 0.7: Structure is basically complete but boundaries are implicit.
③ 0.4: Clause exists but event roles are atypical.
④ 0.1: No relevant clausal structure.

Dimension 3: Event Causal Role Relation Analysis (Wt 0.3)

Analyze the role relationship of the two events within the causal structure:

1. **AgentResult Role:**
   Event 1 is the Subject, Event 2 is the Predicate (or vice versa).
2. **PredicateComplement Role:**
   Event 1 is the Verb, Event 2 is the Object.
3. **Modification Relation:**
   Event 1 modifies Event 2, or Event 2 modifies Event 1.
4. **Parallel Event Configuration:**
   Two events are in a parallel structure (indicates weak causality/concurrency).

Quantitative Scoring Criteria:

① 1.0: Typical causal syntactic configuration.
② 0.7: Reasonable modification or functional dependency.
③ 0.4: Parallel or weak association.
④ 0.1: Syntactic structure does not support causality.

Final Causal Role Heuristic Score: $S_{\text{Causal Role}} = 0.4 \times S_1 + 0.3 \times S_2 + 0.3 \times S_3$

This score characterizes the support strength for causality at the role level, not the final judgment.

Figure 9: The prompt design for causal role heuristic analysis (continued).

LLM Dependency Parsing Prompt Design (Part 1)

Dependency-based Causal Pattern Analysis Instruction

Role and Objective

You are a Dependency-based Causal Pattern Evaluator. Your task is to: based on the given dependency path, identify whether the two event predicates match defined structural causal patterns in the dependency tree, and quantify their causal support strength. Your output should be a continuous confidence score (01) to characterize the degree of support for the candidate causal relationship at the dependency structure level, rather than a final causal judgment.

Figure 10: The prompt design for dependency-based causal pattern analysis.

LLM Dependency Parsing Prompt Design (Part 2)

Dependency-based Causal Pattern Analysis Instruction (Cont.)

Input Format
Sentence: "{sentence}"
Event 1 Predicate: <t1>{pred1}</t1> (POS: {pos1})
Event 2 Predicate: <t3>{pred2}</t3> (POS: {pos2})
Dependency Path: {dependency_path}
(Format Example: pred1 $\xleftarrow{nsubj}$ cause $\xrightarrow{dobj}$ pred2)

Core Dependency Causal Pattern Definitions

Examine the following three typical causal dependency patterns and calculate their matching strength.

Pattern A: Subject-Verb-Object Causal Chain (Wt 0.40)

Structure Prototype:

[Event1 (Subject)] $\xleftarrow{nsubj}$ [Causal Verb] $\xrightarrow{dobj}$ [Event2 (Object)]

Key Criteria:

① Explicit causal verb exists in the path.
② Event 1 is the subject/agent; Event 2 is the object/result.

Causal Verb Strength Grading:

① Strong (1.0): cause, lead to, result in, trigger, induce.
② Medium (0.8): bring about, generate, contribute to, spark.
③ Weak (0.6): affect, relate to, involve, link.
④ Non-causal (0.0): describe, explain, illustrate, express.

Pattern B: Adverbial Clause Causal Chain (Wt 0.35)

Structure Prototype:

[Event1] $\xrightarrow{advcl}$ [Event2] (or reverse)

Allowed Variants:

① Direct advcl: t1 $\xrightarrow{advcl}$ t3
② Explicit Mark: because/since $\xrightarrow{mark}$ t1 + t1 $\xrightarrow{advcl}$ t3
③ Nested Path: Indirect advcl via intermediate nodes.

Key Criteria:

advcl semantics must explicitly indicate Cause/Result/Purpose/Condition. Presence of mark (because/since/if) significantly boosts matching.

Pattern C: Prepositional Causal Phrase Chain (Wt 0.25)

Structure Prototype:

[Event1] $\xleftarrow{nmod/prep}$ [Prep Phrase] $\xrightarrow{pobj}$ [Event2]

Preposition Types & Strength:

① Cause (1.0): because of, due to, owing to, on account of.
② Result (0.8): so that, with the result that, thereby.
③ Purpose (0.6): for, in order to, for the purpose of.
④ Condition (0.5): under condition, on premise.

Key Criteria:

① The prepositional phrase must carry explicit causal semantics.
② Event predicates must serve as core components of the prepositional structure.

Figure 11: The prompt design for dependency-based causal pattern analysis (continued).

LLM Dependency Parsing Prompt Design (Part 3)

Dependency-based Causal Pattern Analysis Instruction (Cont.)

Scoring Calculation Rules

Based on the identified patterns, calculate the confidence scores using the following formulas:

Rule 1: Pattern Match Score (Range: 0-1)

$$Match = 0.7 \times StructureSim + 0.3 \times RoleCorrect$$

① **StructureSim:** Similarity between the extracted dependency path and the target pattern structure $(0-1)$.
② **RoleCorrect:** Rationality of the syntactic roles assumed by the events in the pattern $(0-1)$.

Rule 2: Link Quality (Range: 0-1)

$$LinkQuality = 0.5 \times D + 0.3 \times C + 0.2 \times I$$

1. Directness (D):

$$D = \frac{1}{1 + \max(0, L - L^*)}$$

*Ideal Length L^*: Pattern A=3, Pattern B=2, Pattern C=3.
2. Clarity (C): Causal strength level of the connector/verb $(0-1)$.
3. Integrity (I):
① Complete: 0.8–1.0 ② Partial: 0.5–0.7 ③ Incomplete: < 0.5

Rule 3: Noise Factor Calculation (Range: 0-1)

$$NoiseFactor = 1.0 - \sum Penalty$$

Penalty Items:
① Negation Modifier: -0.4 ② Hypothetical Mood: -0.2
③ Temporal Ambiguity: -0.1 ④ Excessive Distance: $-0.1 \times \max\left(0, \frac{L - L^* - 2}{2}\right)$

Final Integration: Merged Score

Pattern-Specific Adjusted Score ($Score_i$ where $i \in \{A, B, C\}$):

$$Score_i = Match \times LinkQuality \times NoiseFactor \times LexicalStrength$$

*LexicalStrength: The strength of the causal verb/connector defined in the pattern.

Final Dependency Heuristic Score (S_{dep}):

$$S_{dep} = \max(0.40 \cdot Score_A, \quad 0.35 \cdot Score_B, \quad 0.25 \cdot Score_C)$$

The final score reflects the maximum support strength for the event causality.

Figure 12: The prompt design for dependency-based causal pattern analysis (continued).

LLM Temporal-order Prompt Design (Part 1)

Temporal-order-based Causal Heuristic Instruction

Role and Objective
You are a Temporal Relation Evaluator. Your task is to judge whether two events satisfy the basic causal constraint of "Cause precedes Effect" and quantify the support strength of temporal information for causal inference.
Your output should be a continuous score in the [0,1] interval, representing the degree to which the temporal order supports a candidate causal relationship, rather than a final causal judgment.

Input Format
Local Context: "{context}"
Sentence 2: "{sentence2}"
Event 1: <t1>{event1}</t1> Event 2: <t3>{event2}</t3>

Temporal Order Analysis Framework

Please analyze the following temporal phases and calculate the score:

Phase 1: Time Direction Judgment (Core Constraint) (Wt 0.45)

Determine the relative occurrence order of the events:

Scoring Criteria:
① Explicit Sequence (1.0): Clear markers exists, <t1> is explicitly earlier than <t3> (e.g., before, subsequently, then).
② Implicit Sequence (0.6-0.8): Inferred based on context or narrative logic.
③ Simultaneous (0.3-0.5): No obvious sequence distinction.
④ Conflict (0.0): <t3> is explicitly earlier than <t1> (Violates causal constraint).

Record the score for this phase as S_{dir}.

- -

Phase 2: Temporal Expression Consistency (Wt 0.30)

Evaluate if internal temporal information supports the sequence:

Scoring Criteria:
① Strong Support (0.8-1.0): Tense/Aspect indicates completion → subsequent event.
② Neutral (0.5-0.7): Simple tense, relies on external cues.
③ Conflict (0.1-0.3): Contradictory tenses or temporal adverbs.

Record the score for this phase as S_{tense}.

- -

Phase 3: Time Gap Rationality (Wt 0.25)

Judge if the time gap fits a commonsense causal chain:

Scoring Criteria:
① Immediate/Tight (0.8-1.0): Immediately, soon, shortly after.
② Reasonable Gap (0.6-0.8): After a period of time.
③ Excessive/Vague (0.2-0.4): Years later, later on.
④ Unreasonable (0.0-0.2): Extremely long gap without mechanistic explanation.

Record the score for this phase as S_{gap}.

(Continued on next page...)

Figure 13: The prompt design for temporal-order analysis.

LLM Temporal-order Prompt Design (Part 2)

Temporal-order-based Causal Heuristic Instruction (Cont.)

Interference Penalty & Final Score

Penalty Items ($Penalty$): ① Highly Ambiguous Time: -0.1
② Hypothetical/Future Mood: -0.15
③ Interleaved Timelines (Flashback/forward): -0.1

Final Temporal Heuristic Score:

$$h^{temp}(e_i, e_j) = 0.45 \cdot S_{dir} + 0.30 \cdot S_{tense} + 0.25 \cdot S_{gap} - Penalty$$

*Note: This score characterizes the degree of necessity support of temporal information for the causal relationship: Temporal order is a necessary condition for causality, but alone is not sufficient to constitute a sufficient condition.

Figure 14: The prompt design for temporal-order analysis (Continued).

LLM Zero-shot Evaluation Prompt Design

Causality Existence Prompt

Instruction:
You are an expert at identifying cause-effect relationships between events in text, including those that span across sentences. Please answer this question based on the context given to you.
Your final answer must be in JSON format: "Answer":"Yes" or "Answer":"No". Only provide the JSON answer, nothing else.

Input:
text (contains both events): {Document Content} event1: {Event Mention 1} event2: {Event Mention 2}
Question: Is there a causal relationship between {Event Mention 1} and {Event Mention 2} in the text?

Predict:
"Answer": "Yes" (or "No")

Causality Direction Prompt

Instruction:
You are an expert at identifying cause-effect relationships between events in a text. Please answer this question based on the context given to you.
Question: What is the causal relationship between {Event Mention 1} and {Event Mention 2} in the following text?

Options:
PRECONDITION: Event 1 causes/enables Event 2 (Event 1 $\rightarrow$ Event 2).
FALLING_ACTION: Event 2 causes/enables Event 1 (Event 1 $\leftarrow$ Event 2).
NONE: No direct causal relationship.
Your final answer should be in the format: "Answer": "PRECONDITION", "Answer": "FALLING_ACTION", or "Answer": "NONE". Only provide the answer, nothing else.

Input:
text (contains both events): {Document Content} event1: {Event Mention 1} event2: {Event Mention 2}

Predict:
"Answer": "PRECONDITION" (or "FALLING_ACTION", "NONE")

Figure 15: Zero-shot evaluation prompt templates for causality identification and direction classification.