

ICGR: Iterative Causal Graph Refinement for Document-Level Event Causality Identification

Longxuan Ma, Zhuoran Yu, Zhengtao Yu, Shengxiang Gao, Yuxin Huang*, Yantuan Xian

Yunnan Key Laboratory of Artificial Intelligence, Faculty of Information Engineering and Automation, Kunming University of Science and Technology
lxma@kust.edu.cn, yuzhuoran@stu.kust.edu.cn, ztyu@hotmail.com
gaoshengxiang.yn@foxmail.com, huangyuxin2004@163.com, xianyt@kust.edu.cn

Abstract

Document-level event causality identification (DECI) requires identifying causal relations across a document, where inter-sentence causality is often implicit, and supported by weak evidence. Existing methods typically construct event graphs from pairwise predictions, which may overlook weak but valid causal edges, and thus hinder causal message passing. To address these issues, we propose **ICGR**, an **I**terative **C**ausal **G**raph **R**efinement framework that iteratively refines event graphs by preserving low-confidence edges and performing message passing on the refined graph to produce a more complete and reliable structure. In each refinement round, the Constrained Edge Completion module identifies missing causal edges and the Conditional Diffusion module refines low-confidence edges through denoising. Experiments on EventStoryLine and MAVEN-ERE show that ICGR achieves state-of-the-art performance, with notable improvements on inter-sentence causal relations.

1 Introduction

Event Causality Identification (ECI) aims to determine whether one event causes another in text (Berant et al., 2014; Cheng et al., 2025), supporting downstream tasks such as knowledge graph construction (Tao et al., 2023; Huang et al., 2025) and event prediction (Zhan et al., 2024, 2025). Depending on the scope of candidate event pairs, ECI is commonly studied at either the sentence or document level (Mirza and Tonelli, 2014; Gao et al., 2019). Compared with sentence-level ECI (SECI), document-level ECI (DECI) is more challenging because causal relations may span sentences, be expressed implicitly, and be obscured by irrelevant context information (Phu and Nguyen, 2021; Yuan et al., 2023).

To model document-level dependencies, recent DECI methods commonly construct event graphs,

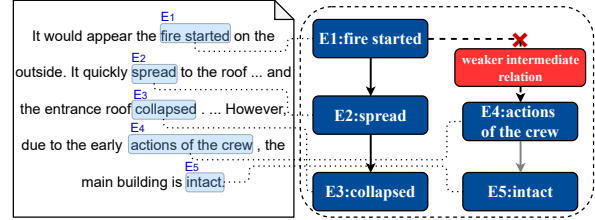


Figure 1: A motivating example where a weak cross-sentence causal link may be missed or assigned low confidence, thereby weakening long-range causal reasoning.

where nodes represent event mentions and directed edges represent causal relations (Cheng et al., 2025). Existing methods mainly follow two paradigms (Liu et al., 2024): **identifying-after-learning**, which first learns event representations and then performs causal identification in a single-step classification paradigm (Yuan et al., 2026), and **identifying-while-learning**, which builds intermediate graphs during prediction (Yuan et al., 2023; Liu et al., 2024). However, both paradigms remain sensitive to the quality of the causal structure during reasoning, where valid edges may be missed or assigned low confidence, limiting reliable message passing for long-range reasoning.

As illustrated in Figure 1, both $E_1 \rightarrow E_2 \rightarrow E_3$ and $E_1 \rightarrow E_4 \rightarrow E_5$ form causal chains, while the latter relies on the less explicit cross-sentence relation $E_1 \rightarrow E_4$. Since graphs are built from pairwise predictions, this underlying relation may either be missed or retained only with low confidence. In both cases, the path $E_1 \rightarrow E_4 \rightarrow E_5$ becomes less effective for supporting subsequent reasoning. This example highlights two issues in prediction constructed event graphs: some valid causal edges may be absent from the graph, while some retained edges may provide only weak confidence support for message passing. These limitations hinder the modeling of causal relation.

To address these issues, we propose **ICGR**, an **I**terative **C**ausal **G**raph **R**efinement framework

** Corresponding author

for DECI. ICGR repeatedly refines document-level causal structures across multiple identification rounds by preserving weak but potentially useful causal edges and completing missing edges. ICGR contains a Constrained Edge Completion module (CECM) and a Conditional Diffusion module (CDM). The CEC model proposes missing edges under confidence constraints, while the CD model provides denoising-based confidence support for weak retained edges. In this way, ICGR treats absent edges and weak retained edges as two complementary refinement targets: the former requires structural completion, whereas the latter requires more stable confidence support before message passing. The refined graph is then used as an intermediate reasoning structure, allowing structural signals to guide event representation learning across refinement rounds. Specifically, the Constrained Edge Completion module proposes missing edges under confidence constraints, restoring potential causal paths that are absent from the initial graph. The Conditional Diffusion module operates on edge-level latent representations and provides denoising-based support for retained edges, making their confidence signals more stable before graph reasoning. Here, diffusion is used as a stability-based confidence support mechanism: it perturbs and denoises weak retained edge representations to check whether they carry stable causal signals before message passing. The refined graph is then used by a confidence-aware graph reasoner to update event representations for the next prediction round. This design has two advantages. First, it separates missing-edge recovery from confidence support, allowing ICGR to complete absent causal links while strengthening weak retained edges before message passing. Second, it avoids the one-shot decision problem of the single-step classification paradigm. The refined graph is used only as an intermediate reasoning structure: it guides confidence-aware message passing and updates event representations across refinement rounds.

Our contributions are summarized as follows:

- We propose ICGR, an iterative causal graph refinement framework for document-level event causality identification. By preserving weak but useful signals during iterative refinement, ICGR better addresses the DECI task.
- ICGR preserves weak but useful causal signals during iterative refinement to improve document-

level causal reasoning. CECM recovers missing causal edges, while CDM provides denoising-based confidence support for retained edges.

- Experiments on EventStoryLine (Caselli and Vossen, 2017) and MAVEN-ERE (Wang et al., 2022) show that ICGR improves document-level causality identification, with clear improvements on inter-sentence causal relations. We give a detailed analysis about the results and our code/data will be released on GitHub¹.

2 Related Work

Event causality identification (ECI) aims to identify causal relations between event mentions in text. Compared with sentence-level ECI, document-level ECI (DECI) is more challenging because causal evidence may be implicit, distributed across sentences, and mixed with irrelevant long-context information (Gao et al., 2019; Phu and Nguyen, 2021; Cheng et al., 2025). To address these challenges, recent studies have moved beyond isolated event-pair classification and incorporated document-level structural reasoning (Zhou and Tang, 2024; Zhou et al., 2025). Early work uses manually designed constraints such as integer linear programming to model document-level causal structures (Gao et al., 2019). Later graph-based methods construct event graphs or event-pair relational graphs to capture structural dependencies and potential causal chains (Phu and Nguyen, 2021; Chen et al., 2022; Wang et al., 2024). Recent work also explores multi-level encoder architectures to encode event information at sentence, event, event-pair, and discourse levels, further showing the importance of context-aware document modeling for ECI (Chen et al., 2025; Liang et al., 2026). Other studies further improve long-range reasoning by exploiting easier intra-sentence predictions (Xu et al., 2022), learning sparse event representations for intra- and inter-sentence reasoning (Yuan et al., 2023), updating event representations with causal structures (Liu et al., 2024), or modeling high-order dependencies with hypergraphs (Xiang et al., 2025). LLMs have also been explored for ECI, but remain prone to hallucinated or false-positive causal predictions in fine-grained settings (Gao et al., 2023; Cai et al., 2025; Cheng et al., 2025).

Despite these advances, graph-based methods still rely on the quality of intermediate causal

¹<https://github.com/malongxuan/ICGR>

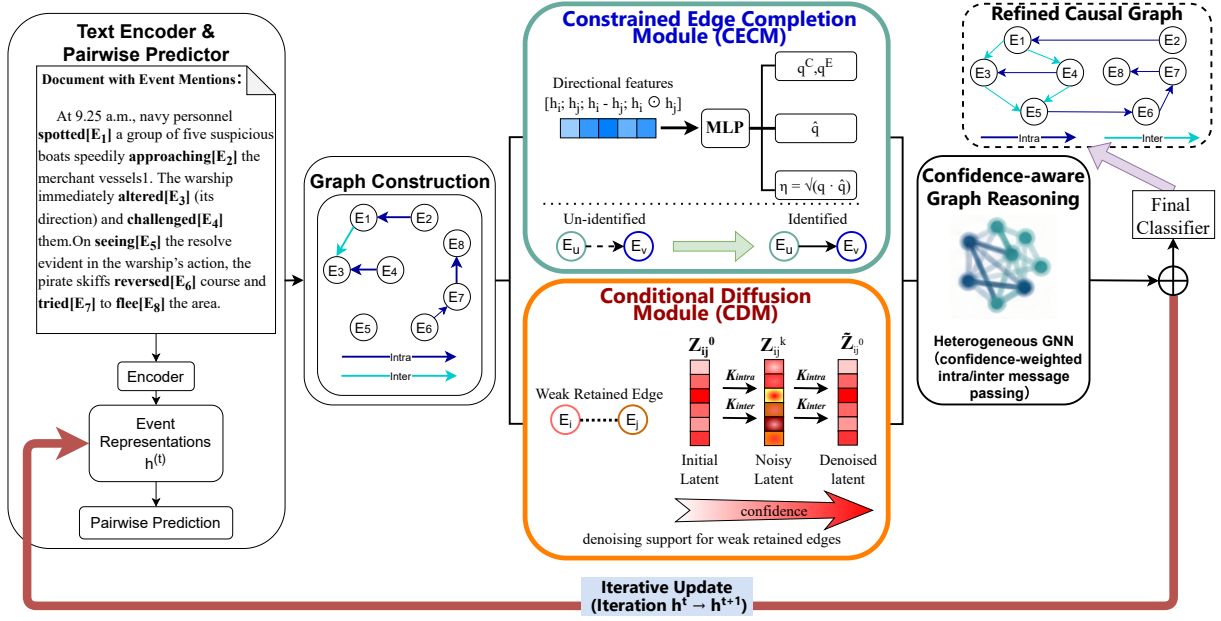


Figure 2: Overview of ICGR. Each round constructs a sparse directed event graph from current predictions, refines missing and weak edges, and updates event representations through confidence-aware heterogeneous graph reasoning.

graphs. When implicit inter-sentence edges are missed or assigned low confidence, message passing may fail to propagate useful evidence. Moreover, most methods focus on whether a causal relation exists but neglect its direction (Liu et al., 2024). Diffusion-based methods such as DiffusECI (Man et al., 2024) denoise event-pair representations, but mainly for pair-level prediction. *ICGR instead uses diffusion within iterative graph refinement to support weak retained edges before message passing, making confidence-conditioned graph refinement its key distinction.*

3 Methodology

Given a document D with event mentions $E = \{e_i\}_{i=1}^n$, document-level event causality identification aims to predict the causal label for each candidate event pair $(e_i, e_j) \in \mathcal{P}$. We use a fixed canonical pair order and define $y_{ij} \in \{\text{NONE}, \text{CAUSE}, \text{EFFECT}\}$, where CAUSE denotes $e_i \rightarrow e_j$ and EFFECT denotes $e_j \rightarrow e_i$. Each pair is associated with a relation type $r(i, j) \in \{\text{intra}, \text{inter}\}$ according to whether the two events occur in the same sentence.

3.1 Framework Overview

Figure 2 summarizes one refinement round of ICGR. Starting from the current event representations $\{h_i^{(t)}\}$, the pairwise predictor first estimates causal labels for all candidate event pairs

and constructs a conservative sparse graph from high-confidence non-NONE predictions. Candidate pairs not retained by this rule remain absent from the current graph.

ICGR then handles retained edges and graph-absent candidates through two complementary refinement mechanisms. The **Constrained Edge Completion Module (CECM)** calibrates retained edges and evaluates graph-absent candidate pairs for constrained edge completion. The **Conditional Diffusion Module (CDM)**, in contrast, provides denoising-based confidence support only for retained edges during graph refinement, without changing their causal directions or introducing new edges. The resulting refined graph is subsequently used by a confidence-aware graph reasoner to update the event representations from $\{h_i^{(t)}\}$ to $\{h_i^{(t+1)}\}$. These updated representations are reused by the pairwise predictor in the next refinement round, and the final labels are produced by the pairwise classifier after adaptive stopping.

3.2 Context Encoding and Pairwise Prediction

We use a pre-trained Transformer encoder to obtain contextualized token representations. For each event mention e_i , we average the token embeddings within its trigger span to obtain $\mathbf{x}_i \in \mathbb{R}^d$, and project it into the graph reasoning space:

$$\mathbf{h}_i^{(0)} = \mathbf{W}_p \mathbf{x}_i. \quad (1)$$

At refinement round t , for each candidate pair (e_i, e_j) , we construct a pair representation and predict its causal label distribution:

$$\begin{aligned} \mathbf{r}_{ij}^{(t)} &= [\mathbf{x}_i; \mathbf{x}_j; \mathbf{h}_i^{(t)} - \mathbf{h}_j^{(t)}], \\ \mathbf{p}_{ij}^{(t)} &= \text{softmax} \left(\text{MLP}_{cls}(\mathbf{r}_{ij}^{(t)}) \right), \end{aligned} \quad (2)$$

where $\mathbf{p}_{ij}^{(t)} = [p_{ij}^{N,(t)}, p_{ij}^{C,(t)}, p_{ij}^{E,(t)}]$ denotes the probabilities of NONE, CAUSE, and EFFECT, respectively.

3.3 Conservative Sparse Graph Construction

At each refinement round, we construct a sparse directed event graph from the current pairwise predictions. The graph contains one node type, *event*, and two directed edge types, *intra* and *inter*. The edge type is determined by whether the two events appear in the same sentence.

For each candidate pair (e_i, e_j) , the predicted label, confidence, and induced directed edge are:

$$\begin{aligned} \hat{y}_{ij}^{(t)} &= \arg \max_c p_{ij,c}^{(t)}, & \hat{p}_{ij}^{(t)} &= \max_c p_{ij,c}^{(t)}, \\ d_{ij}^{(t)} &= \begin{cases} e_i \rightarrow e_j, & \hat{y}_{ij}^{(t)} = \text{CAUSE}, \\ e_j \rightarrow e_i, & \hat{y}_{ij}^{(t)} = \text{EFFECT}, \\ \emptyset, & \hat{y}_{ij}^{(t)} = \text{NONE}, \end{cases} \quad (3) \\ d_{ij}^{(t)} &\text{ is retained if } d_{ij}^{(t)} \neq \emptyset \text{ and } \hat{p}_{ij}^{(t)} > \omega. \end{aligned}$$

During both training and inference, the graph is induced from model predictions rather than gold causal edges. Gold labels are used only to compute supervised losses. The threshold ω controls the precision–coverage trade-off of the initial topology. A lower value may introduce noisy edges, while an overly strict value may remove useful weak causal links before refinement. Thus, the constructed graph is a conservative structural backbone rather than a complete causal graph. Self-loops are added to preserve node identity during graph reasoning. Since edge selection is discrete, gradients are not propagated through graph construction; the classifier, refinement modules, and graph reasoner are trained through their differentiable objectives.

3.4 Constrained Edge Completion Module

The conservative sparse graph may miss valid causal edges, especially implicit inter-sentence relations. We introduce the **Constrained Edge Completion Module** (CECM) to address two complementary graph-refinement needs: calibrating retained edges and recovering reliable missing edges.

CECM contains a directional classification head and a direction-agnostic existence head. Its outputs are used only to refine the intermediate graph before message passing, rather than as a separate final decoder.

For a candidate pair (e_i, e_j) at round t , CECM first constructs a graph-space directional representation:

$$\begin{aligned} \mathbf{u}_{ij}^{(t)} &= [\mathbf{h}_i^{(t)}; \mathbf{h}_j^{(t)}; \mathbf{h}_i^{(t)} - \mathbf{h}_j^{(t)}; \mathbf{h}_i^{(t)} \odot \mathbf{h}_j^{(t)}], \\ \mathbf{s}_{ij}^{(t)} &= \text{MLP}_{enc}(\mathbf{u}_{ij}^{(t)}). \end{aligned} \quad (4a)$$

From the shared representation $\mathbf{s}_{ij}^{(t)}$, the two prediction heads compute:

$$\mathbf{q}_{ij}^{(t)} = \text{softmax} \left(\mathbf{W}_{disc} \mathbf{s}_{ij}^{(t)} + \mathbf{b}_{disc} \right), \quad (4b)$$

$$\hat{q}_{ij}^{(t)} = \sigma \left(\mathbf{w}_{conf}^\top \mathbf{s}_{ij}^{(t)} + b_{conf} \right). \quad (4c)$$

Here, $\mathbf{q}_{ij}^{(t)} = [q_{ij}^{N,(t)}, q_{ij}^{C,(t)}, q_{ij}^{E,(t)}]$ is a three-way directional distribution: $q_{ij}^{C,(t)}$ supports $e_i \rightarrow e_j$, whereas $q_{ij}^{E,(t)}$ supports $e_j \rightarrow e_i$ under the canonical pair order. In contrast, $\hat{q}_{ij}^{(t)}$ is a scalar direction-agnostic existence score indicating whether the pair contains any causal relation. Thus, the two heads provide complementary evidence: $\mathbf{q}_{ij}^{(t)}$ specifies which causal direction is supported, whereas $\hat{q}_{ij}^{(t)}$ estimates whether a causal edge exists.

To make the existence head directly affect graph refinement, we combine directional and existence confidence:

$$\eta_{ij}^{C,(t)} = \sqrt{q_{ij}^{C,(t)} \hat{q}_{ij}^{(t)}}, \quad \eta_{ij}^{E,(t)} = \sqrt{q_{ij}^{E,(t)} \hat{q}_{ij}^{(t)}}. \quad (5)$$

The resulting η is an existence-aware directional confidence. It is high only when both causal existence and the corresponding direction receive strong support. This design allows the existence head to directly participate in graph refinement without introducing an additional threshold.

During training, CECM is supervised by the pairwise causal label:

$$\begin{aligned} \mathcal{L}_{disc}^{(t)} &= \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \left[\text{CE}(\mathbf{q}_{ij}^{(t)}, y_{ij}) \right. \\ &\quad \left. + \lambda_{conf} (\hat{q}_{ij}^{(t)} - \mathbf{1}[y_{ij} \neq \text{NONE}])^2 \right]. \end{aligned} \quad (6)$$

Retained-edge calibration. For a retained classifier-induced edge, we fuse the main classifier confidence with the existence-aware directional

confidence. If $d_{ij}^{(t)} = e_i \rightarrow e_j$, the calibrated edge confidence is:

$$\tilde{a}_{d_{ij}}^{(t)} = (1 - \alpha_{disc})p_{ij}^{C,(t)} + \alpha_{disc}\eta_{ij}^{C,(t)}. \quad (7)$$

If $d_{ij}^{(t)} = e_j \rightarrow e_i$, it is:

$$\tilde{a}_{d_{ij}}^{(t)} = (1 - \alpha_{disc})p_{ij}^{E,(t)} + \alpha_{disc}\eta_{ij}^{E,(t)}. \quad (8)$$

Missing-edge completion. CECM also evaluates candidate pairs that are absent from the current sparse graph, including pairs predicted as NONE by the main pairwise classifier. For graph augmentation, CECM is applied to ordered event inputs to score candidate directed edges. A proposal $e_u \rightarrow e_v$ is activated when the ordered existence-aware confidence $\eta_{uv}^{C,(t)}$ exceeds τ_{disc} and the proposal satisfies two structural constraints: a sentence-distance window W and a per-source top- k outgoing constraint. These constraints prevent uncontrolled graph densification while allowing reliable missing edges to participate in message passing.

For each newly completed inter-sentence edge, we compute a reverse-order consistency confidence:

$$a_{uv}^{comp,(t)} = \frac{1}{2} \left(\eta_{uv}^{C,(t)} + \eta_{vu}^{E,(t)} \right). \quad (9)$$

Under the reverse ordered input (e_v, e_u) , the EFECT label also corresponds to the same direction $e_u \rightarrow e_v$. Thus, Eq. 9 favors proposed edges that are directionally stable under both input orders and are supported by edge-existence confidence. The completed edge is used only as an intermediate message-passing edge. Ordered scoring does not create additional supervised instances for the main classifier.

3.5 Conditional Diffusion Module

CECM mainly improves graph completeness, while some retained edges may still have weak or unstable confidence. Such edges are important in document-level ECI: removing them too early may break long-range causal chains, whereas directly propagating their raw confidence may introduce noise. We therefore introduce the **Conditional Diffusion Module** (CDM) as a stability-based confidence support mechanism for retained edges. CDM perturbs edge-level latent representations and denoises them under relation-aware conditions to derive an additional support signal before message

passing. During graph refinement, CDM has a restricted role: its input is a retained edge, its output is an additional confidence-support signal, and it neither activates graph-absent edges nor changes causal directions.

Using the directional representation $\mathbf{u}_{ij}^{(t)}$ in Eq. 4a, we obtain a clean edge latent:

$$\mathbf{z}_{ij}^{0,(t)} = \text{MLP}_z(\mathbf{u}_{ij}^{(t)}). \quad (10)$$

We apply a standard forward noising process:

$$\mathbf{z}_{ij}^{k,(t)} = \sqrt{\bar{\alpha}_k} \mathbf{z}_{ij}^{0,(t)} + \sqrt{1 - \bar{\alpha}_k} \boldsymbol{\epsilon}, \quad (11)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

A relation-aware denoiser predicts the injected noise conditioned on the noised latent, the clean latent, and the timestep:

$$\hat{\boldsymbol{\epsilon}}_{ij} = \epsilon_{\theta}^{r(i,j)} \left(\mathbf{z}_{ij}^{k,(t)}, \mathbf{z}_{ij}^{0,(t)}, k \right). \quad (12)$$

The diffusion loss is:

$$\mathcal{L}_{diff}^{(t)} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \|\boldsymbol{\epsilon}_{ij} - \hat{\boldsymbol{\epsilon}}_{ij}\|_2^2. \quad (13)$$

During refinement, deterministic reverse denoising yields $\tilde{\mathbf{z}}_{ij}^{0,(t)}$. The denoised latent is used to compute a support gate:

$$g_{ij}^{(t)} = \sigma \left(\text{MLP}_g \left[\mathbf{z}_{ij}^{0,(t)}; \tilde{\mathbf{z}}_{ij}^{0,(t)} \right] \right),$$

$$a_{d_{ij}}^{(t)} = \text{clip} \left(\tilde{a}_{d_{ij}}^{(t)} + \lambda_{r(i,j)} g_{ij}^{(t)}, 0, 1 \right). \quad (14)$$

The resulting gate provides bounded additional support to the retained-edge confidence before message passing.

3.6 Confidence-Aware Graph Reasoning

Given the refined graph and its edge confidence scores, we update event representations with a heterogeneous Causal Graph Enhancement (CGE) network. The graph topology is determined by conservative graph construction and constrained completion, while the confidence signals entering message passing are calibrated and supported by CECM and CDM. CGE treats intra-sentence and inter-sentence edges as distinct relation types and performs confidence-aware multi-head message passing.

For an edge $u \rightarrow v$ of relation type r , attention head m , and graph layer ℓ , we compute:

$$e_{uv}^{r,m,\ell} = \phi_r^{m,\ell}(\mathbf{h}_u^{\ell-1}, \mathbf{h}_v^{\ell-1}) + \psi_r^{m,\ell}([\mathbf{h}_u^{\ell-1}; \mathbf{h}_v^{\ell-1}])$$

$$+ \mu a_{uv}^{(t)} + b_r^{m,\ell}. \quad (15)$$

Here, $\phi_r^{m,\ell}(\cdot)$ is a relation-specific source-target compatibility function, $\psi_r^{m,\ell}(\cdot)$ is a learnable pairwise edge-bias function, $a_{uv}^{(t)}$ is the refined edge confidence, μ controls the confidence bias, and $b_r^{m,\ell}$ is a relation-specific bias.

The normalized attention coefficient and relation-specific message are:

$$\begin{aligned}\alpha_{uv}^{r,m,\ell} &= \frac{\exp(e_{uv}^{r,m,\ell}/\tau)}{\sum_{u' \in \mathcal{N}_r(v)} \exp(e_{u'v}^{r,m,\ell}/\tau)}, \\ \mathbf{m}_v^{r,m,\ell} &= \sum_{u \in \mathcal{N}_r(v)} \alpha_{uv}^{r,m,\ell} \mathbf{w}_r^{m,\ell} \mathbf{h}_u^{\ell-1}.\end{aligned}\quad (16)$$

Messages from multiple heads are combined to obtain $\mathbf{m}_v^{r,\ell}$ for each relation type.

Messages from intra-sentence and inter-sentence relation types are fused as:

$$\mathbf{m}_v^\ell = \beta_{intra} \mathbf{m}_v^{intra,\ell} + \beta_{inter} \mathbf{m}_v^{inter,\ell}. \quad (17)$$

If a node receives messages from only one relation type, the available relation-specific message is directly used. We then apply residual connections, layer normalization, and a feed-forward block to stabilize graph reasoning. When multiple graph layers or hops are used, their outputs are fused with Jumping Knowledge. A document-level attention readout is injected back into event representations through a gated residual update, yielding $\mathbf{h}_v^{(t+1)}$ for the next refinement round.

3.7 Iterative Refinement and Training Objective

ICGR performs bounded iterative refinement. The model runs for at least $N_{\min}$ refinement rounds and stops when either the maximum refinement budget is reached or the structural difference between two consecutive predictions is no larger than a threshold δ .

Let $\hat{\mathbf{y}}^{(t)}$ denote the vector of predicted labels at round t . The structural change and stopping criterion are:

$$\Delta^{(t)} = \sum_{(i,j) \in \mathcal{P}} \mathbf{1}[\hat{y}_{ij}^{(t)} \neq \hat{y}_{ij}^{(t-1)}], \quad (18)$$

$$\text{stop} \iff t \geq N_{\min} \text{ and } \Delta^{(t)} \leq \delta.$$

Refinement also stops when the maximum refinement budget is reached. The actual number of executed rounds for document D is denoted as T_D .

At each refinement round, the main classifier is trained with focal loss:

$$\mathcal{L}_{cls}^{(t)} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \mathcal{L}_{focal}(\mathbf{p}_{ij}^{(t)}, y_{ij}). \quad (19)$$

The full training objective is:

$$\mathcal{L} = \sum_{t=0}^{T_D-1} \left[\frac{\mathcal{L}_{cls}^{(t)}}{t+1} + \lambda_{disc} \mathcal{L}_{disc}^{(t)} + \lambda_{diff} \mathcal{L}_{diff}^{(t)} \right]. \quad (20)$$

The final causal prediction is produced by the pairwise classifier at the last executed refinement round:

$$\hat{y}_{ij} = \arg \max_c p_{ij,c}^{(T_D-1)}. \quad (21)$$

The refined graph therefore serves only as an intermediate reasoning structure. Across refinement rounds, CECM and CDM modify the topology or confidence used for message passing, the graph reasoner updates the event representations, and these updated representations affect the pairwise predictions in the next round. The intermediate graph edges themselves are not directly used as final causal labels.

Further methodological details are provided in Appendix A.

4 Experiments

4.1 Datasets, Baselines, and Evaluation Metrics

We evaluate ICGR on two benchmarks: **EventStoryLine (v0.9)** (Caselli and Vossen, 2017) and **MAVEN-ERE** (Wang et al., 2022). Detailed dataset statistics, preprocessing protocols, and split settings are described in Appendix D. We compare ICGR with a comprehensive set of baselines, ranging from pre-trained language models to state-of-the-art graph-based and generative methods. Detailed descriptions of these baselines are provided in Appendix C, while the settings and details of the LLM-based comparisons are presented in Appendix G. We additionally report other LLM results that are only available under the intra-sentence setting in Appendix G.3. Implementation details and equation-level parameter settings are provided in Appendix B. Following previous work (Phu and Nguyen, 2021), we use precision (P), recall (R), and F1 as evaluation metrics. We report results on **Intra-sentence**, **Inter-sentence**, and **Overall** settings to assess the model’s ability to capture long-distance causal dependencies. The label configurations for causality direction identification and causality existence identification are detailed in Appendix H.

Model	Intra-sentence			Inter-sentence			Intra+Inter (Overall)		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
<i>Causality Direction Identification (Ternary Classification)</i>									
BERT (Devlin et al., 2018)	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
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 (Liu et al., 2024)	17.5	17.0	17.2	6.8	19.2	10.0	8.3	18.5	11.5
LLaMA-3.1-8B(zero-shot)	11.5	92.0	20.4	0.9	92.6	1.8	2.6	92.2	5.1
ICGR (Ours)	70.2	55.8	62.2	45.6	45.3	45.4	52.4	48.3	50.3
<i>Causality Existence Identification (Binary Classification)</i>									
BERT (Devlin et al., 2018)	60.4	45.7	52.0	30.6	39.1	34.3	37.2	41.2	39.1
RoBERTa (Liu et al., 2019)	62.7	45.4	52.7	32.7	38.3	35.3	39.7	40.6	40.1
ERGO (Chen et al., 2022)	49.7	72.6	59.0	43.2	48.8	45.8	46.3	50.1	48.1
SENDIR (Yuan et al., 2023)	65.8	66.7	66.2	33.9	90.0	48.3	37.8	82.8	51.9
iLIF (Liu et al., 2024)	78.6	66.3	71.2	53.5	65.9	59.1	59.2	66.1	62.5
DiffusECI (Man et al., 2024)	65.3	78.3	71.4	61.9	59.9	60.9	63.0	64.1	63.5
NCHM (Xiang et al., 2025)	74.9	65.8	69.5	59.6	52.6	55.6	64.6	56.7	60.0
MEA (Liang et al., 2026)	60.1	71.7	65.4	50.4	63.7	56.3	51.2	69.6	59.0
LLaMA-2-7B (Liu et al., 2024)	26.9	29.3	28.0	10.8	31.9	16.1	13.2	31.1	18.5
LLaMA-3.1-8B(zero-shot)	19.1	95.6	31.8	1.6	96.2	3.1	4.5	95.8	8.5
LLaMA-3.1-8B(few-shot)	22.2	95.9	36.1	1.9	96.4	3.7	5.2	96.1	9.9
ICGR (Ours)	80.8	67.6	73.6	64.3	60.2	62.2	65.0	64.8	65.0

Table 1: Experimental results on the **EventStoryLine (v0.9)** dataset. Results are reported as the average over 5 folds. The best results are marked in **bold**.

4.2 Results and Analysis

Tables 1 and 2 report the main results. ICGR achieves the best overall F1 on both datasets under both direction and existence identification, showing the effectiveness of iterative causal graph refinement.

The gains are especially clear for inter-sentence causality, where relations are often implicit and sensitive to early prediction errors. By completing missing edges and reinforcing weak retained edges, ICGR provides more useful structure for long-distance reasoning. The direction-identification gains further suggest that refinement preserves directional information rather than merely adding causal edges. Significance analysis under the EventStoryLine five-fold protocol is reported in Appendix F.

4.3 Ablation Study

Table 3 reports five-fold ablation results on EventStoryLine (v0.9). Removed modules are replaced with neutral inputs of the same shape. Ablation results on MAVEN-ERE are reported in Appendix D.3.

w/o CDM. Removing the Conditional Diffusion module (CDM) consistently degrades performance under both direction identification and existence

identification. This indicates that denoising edge-level latent representations provides useful support for weak retained edges, especially when pairwise evidence alone is not sufficiently stable.

w/o CECM. Removing the document-level Constrained Edge Completion module (CECM) also causes a clear degradation. This suggests that the graphs constructed from predictions still miss part of the useful directional structure and that constrained causal edge completion provides complementary topological cues beyond local pairwise decoding.

w/o Iteration. Removing iterative refinement leads to the largest overall degradation among all variants. This shows that iterative refinement is crucial for progressively correcting sparse graph structure and propagating causal evidence across the document.

Overall, the ablations support the complementary roles of CDM, CECM, and iterative refinement. Controlled variants further show that the gains are not merely due to additional prediction capacity or stochastic regularization: replacing CDM with a parameter-matched MLP gate or a noise-augmented gate yields 47.7/61.4 and 48.6/62.9 overall F1 for direction/existence, respectively, while using CECM only as a prediction-level

Model	Intra-sentence			Inter-sentence			Intra+Inter (Overall)		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
<i>Causality Direction Identification (Ternary Classification)</i>									
BERT (Devlin et al., 2018)	46.0	48.8	47.4	42.1	45.2	43.6	42.4	45.5	43.9
ERGO (Chen et al., 2022)	62.3	63.1	62.7	47.8	59.8	53.1	48.7	60.1	53.8
SENDIR (Yuan et al., 2023)	46.6	44.2	45.4	46.8	43.0	44.8	46.8	43.1	44.9
iLIF (Liu et al., 2024)	73.7	49.7	59.4	66.3	47.5	55.3	66.9	47.6	55.6
ICGR (Ours)	73.3	52.5	61.2	65.9	50.2	57.0	65.8	51.3	57.7
<i>Causality Existence Identification (Binary Classification)</i>									
BERT (Devlin et al., 2018)	46.8	50.3	48.5	43.0	46.8	44.8	43.3	47.1	45.1
ERGO (Chen et al., 2022)	63.1	65.3	64.2	48.7	62.0	54.6	49.6	62.3	55.2
SENDIR (Yuan et al., 2023)	51.4	53.6	52.5	51.9	52.8	52.4	51.9	52.9	52.4
iLIF (Liu et al., 2024)	74.4	51.5	60.9	67.1	49.2	56.8	67.7	49.4	57.1
GPT-3.5-Turbo (Gao et al., 2023)	19.9	85.8	32.3	-	-	-	-	-	-
GPT-4 (Gao et al., 2023)	22.5	92.4	36.2	-	-	-	-	-	-
ICGR (Ours)	75.5	53.9	62.9	67.0	52.3	58.7	67.5	52.9	59.3

Table 2: Experimental results on the MAVEN-ERE. Results are reported as the average over five random seeds.

Model	Direction			Existence		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
ICGR (Full Model)	52.4	48.3	50.3	65.0	64.8	65.0
w/o CDM	49.5	42.6	45.7	59.4	58.2	58.8
w/o CECM	50.1	44.1	46.8	60.8	60.3	60.5
w/o Iteration	43.0	40.9	41.9	45.4	45.8	45.6

Table 3: Ablation results on EventStoryLine (v0.9).

ensemble yields 48.6/61.9, all below the full ICGR (50.3/65.0). Detailed settings are provided in Appendix A.7.

4.4 Parameter Sensitivity Analysis

We analyze four key hyperparameters that control different stages of ICGR: the relation-specific CDM denoising depth, the CECM completion threshold τ_{disc} , the number of refinement iterations, and the sparse graph retention threshold ω . Together, these analyses examine the effects of confidence support, edge completion, iterative reasoning depth, and initial graph sparsity.

Sensitivity to CDM Denoising Depth. Figure 3 shows that intra- and inter-sentence relations prefer different denoising depths. Intra-sentence relations favor a shallower schedule, suggesting that local causal evidence is relatively concentrated and may be over-smoothed by excessive denoising. In contrast, inter-sentence relations benefit from deeper denoising, indicating that weak long-distance dependencies require stronger refinement to obtain useful confidence support. This result supports relation-aware denoising rather than a uniform schedule for all causal relations.

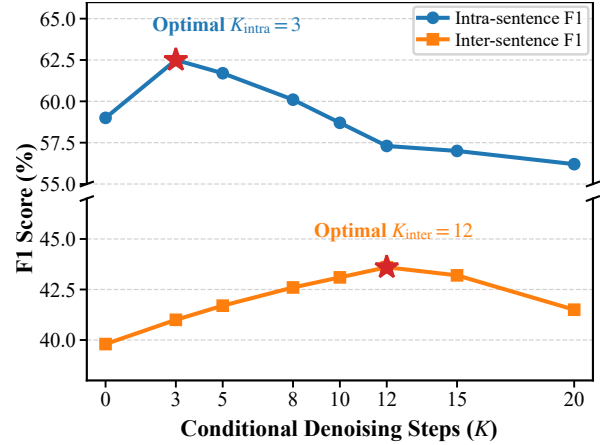


Figure 3: Impact of CDM denoising sampling steps on intra- and inter-sentence causality.

Sensitivity to Edge Completion Threshold. Figure 4 shows the effect of the CECM completion threshold τ_{disc} applied to the existence-aware directional confidence η . The results reveal a precision-coverage trade-off: lower thresholds increase coverage but may activate noisy causal proposals, whereas overly strict thresholds miss useful completed edges. We therefore adopt a conservative high-threshold setting. Together with the sentence-distance window and per-source top- k constraint, this setting limits graph densification and allows only a small number of high-confidence, directionally reliable candidate edges to participate in message passing.

Sensitivity to Refinement Iterations. Figure 5 shows that performance improves during the early refinement rounds but decreases when the number of iterations becomes too large. Moderate iterative reasoning allows causal information to propagate

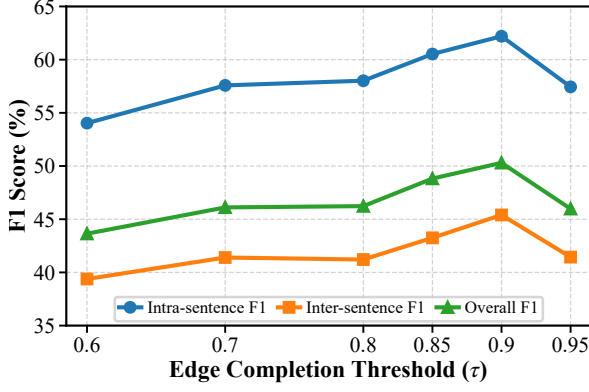


Figure 4: Impact of the CECM completion threshold τ_{disc} on edge completion.

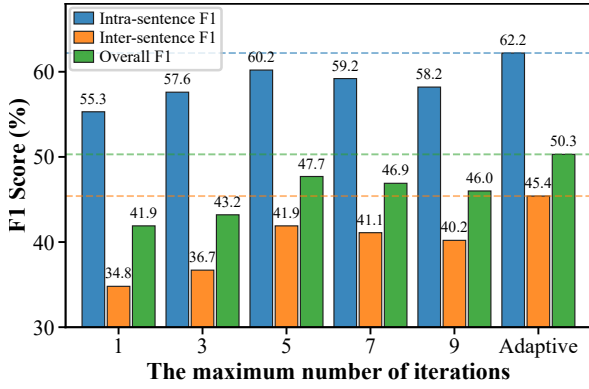


Figure 5: Impact of the number of refinement iterations.

over a broader structural context, whereas excessive refinement may lead to over-smoothing and accumulated message-passing noise. The adaptive stopping strategy achieves the best overall performance, supporting bounded rather than unrestricted iterative refinement.

Sensitivity to Sparse Graph Retention Threshold. Figure 6 reports the effect of the sparse graph retention threshold ω . Lower thresholds retain more classifier-induced edges but also introduce noisy relations, whereas higher thresholds may prune weak but useful causal links before refinement. The best performance is obtained at an intermediate threshold, supporting the use of a conservative but not overly sparse initial structural backbone.

Interpretation of Hyperparameters. The four sensitivity analyses correspond to distinct and interpretable stages of ICGR rather than independent tuning objectives. The retention threshold ω controls the sparsity of the initial graph, while τ_{disc} controls the precision–coverage trade-off of subsequent edge completion. The relation-specific denoising depths determine how strongly retained

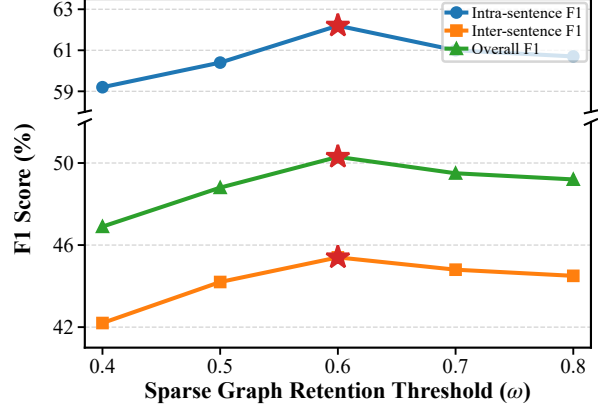


Figure 6: Impact of the sparse graph retention threshold ω on graph construction.

edges are refined before message passing, and the refinement budget controls the depth of iterative structural reasoning. The sentence-distance window W and per-source top- k constraint serve as additional safeguards against uncontrolled graph densification.

4.5 Exploration Study

We further investigate how the causal graph evolves during iterative refinement. Qualitative analysis suggests that the proposed framework gradually improves graph quality in two complementary ways: weak long-distance causal links receive stronger support from reinforced retained edges, while some spurious directions are corrected after repeated topology-aware reasoning. Detailed visualizations and case analyses are provided in Appendix E.

5 Conclusion

We present **ICGR**, a confidence-conditioned structural refinement framework for document-level event causality identification. ICGR alternates between causal prediction over directed event pairs and confidence-aware graph reasoning, refining sparse causal structures through the Conditional Diffusion module, the Constrained Edge Completion module, and heterogeneous graph reasoning. By separating conservative sparse graph construction from subsequent confidence-conditioned refinement, ICGR strengthens weakly supported retained edges and proposes a small number of high-confidence missing causal edges during iterative reasoning. Experiments on EventStoryLine and MAVEN-ERE show that ICGR achieves state-of-the-art performance in both causality direction identification and causality existence identification, especially in overall document-level evaluation.

Limitations

Despite the promising results demonstrated in this study, we acknowledge several limitations regarding evaluation scope.

Regarding the evaluation scope of Large Language Models (LLMs), the evaluation coverage of LLM baselines is not fully uniform. We additionally evaluate LLaMA-3.1-8B under intra-sentence, inter-sentence, and overall settings on EventStoryLine, whereas Qwen3-Max is evaluated only under the intra-sentence setting due to the prohibitive cost of exhaustive pairwise reasoning over long contexts. Some other LLM results are cited from prior work when available. Accordingly, LLM comparisons involving missing inter-sentence or overall results should be interpreted as partial reference comparisons rather than fully controlled document-level comparisons. Therefore, these LLM results are used to illustrate the current limitations of prompting-based LLMs for fine-grained causal discrimination, rather than to claim a complete document-level comparison against all LLM baselines.

Ethics Statement

This work strictly adheres to the ACL Ethics Policy. The datasets utilized in this paper, EventStoryLine and MAVEN-ERE, are publicly available benchmarks widely adopted within the research community. We have not collected any private data or personally identifiable information. However, it is important to note that, consistent with most Natural Language Processing systems trained on real-world text, the proposed model may reflect biases inherent in the underlying data. If deployed in downstream applications, such biases could influence the determination of causal relations. Furthermore, the model must not be used to fabricate misleading causal narratives or disseminate disinformation. During the preparation of this work, we used DeepSeek² and ChatGPT³ solely for language polishing and grammar checking. All technical content, experimental design, and conclusions were produced and verified by the authors.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of

China (U23A2038, 62566033); Major Science Technology Special Projects of Yunnan Province (202302AD080003, 202303AP140008, 202402AG050007, 202402AD080002, 202502AD080016, 202602AD080015), Yunnan Fundamental Research Projects (202301AS070047, 202301AT070471, 202601CF070076, 202601AS070065).

References

- Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1499–1510.
- Ruichu Cai, Shengyin Yu, Jiahao Zhang, and 1 others. 2025. Dr. eci: Infusing large language models with causal knowledge for decomposed reasoning in event causality identification. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 9346–9375.
- 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.
- Chao Chen, Pengfei Luo, Changkai Feng, Tian Wu, Wenbin Jiang, and Tong Xu. 2025. Maps: A multi-task framework with anchor point sampling for zero-shot entity linking. *Data Intell.*, 7(4):1085–1107.
- Meiqi Chen and 1 others. 2022. Ergo: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2025. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*, 58(3):Article 59.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- J. Gao, X. Ding, B. Qin, and 1 others. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. *arXiv preprint arXiv:2305.07375*.
- 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*:

²<https://www.deepseek.com/>

³<https://chatgpt.com/>

- Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817.
- Xiang Gao, Weiqing Chen, Ying Cui, Xiang Dai, and Lican Dai. 2025. Progressive adversarial contrastive learning: Towards efficient data augmentation in adversarial defense. *Data Intelligence*, 7(4):1169–1191.
- Feijuan He, Han Lai, Jiaqi Liu, Bo Wang, Haoran Chen, Haohan Liu, and Chenxi Zhang. 2025. Solving mathematical problems using large language models: A survey. *Data Intelligence*, 7(4):907–946.
- Zhisheng Huang, Xudong Jia, Tao Chen, and Zhongwei Zhang. 2025. A general scalable approach for knowledge selection based on iterative hybrid encoding and re-ranking. *Data Intelligence*, 7(1):124–142.
- Hao Liang, Qifeng Zhou, Wanyuan Gong, and Xiang Li. 2026. Multi-level encoder architectures for event causality identification. *Neural Networks*, page 108356.
- C. 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 (ACL)*.
- Yinhan Liu, Myle Ott, Naman Goyal, and 1 others. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Mastering context-to-label representation transformation for event causality identification with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18760–18768.
- Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106.
- 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.
- 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.
- Jing Wang, Shuo Zhang, and Runzhi Li. 2024. Gate feature interaction network for relation prediction in knowledge graph. *Data Intell.*, 6(3):749–770.
- Xiaozhi Wang, Yulin Chen, Ning Ding, and 1 others. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.
- Wei Xiang, C. Liu, and Bang Wang. 2025. Modeling document causal structure with a hypergraph for event causality identification. *Neural Networks*, 184:107080.
- Zhenran Xu, Baotian Hu, and 1 others. 2022. Ppat: Progressive graph pairwise attention network for event causality identification.
- Changsen Yuan, Heyan Huang, Yixin Cao, and 1 others. 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.
- Jianbo Yuan, Yitai Xu, Wen Zhou, Miao Yu, Xiaomin Sun, Liangji Zhang, and Jiayi Luo. 2026. Cognitive-driven hierarchical reasoning for event causality identification. *Knowledge-Based Systems*, page 115914.
- Chuanhong Zhan, Wei Xiang, Liang Chao, and Bang Wang. 2024. What would happen next? predicting consequences from an event causality graph. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 820–832.
- Chuanhong Zhan, Wei Xiang, Chao Liang, and Bang Wang. 2025. Unified semantic and structural modeling via focal attention for causality graph event prediction. *IEEE Transactions on Audio, Speech and Language Processing*.
- Jingshi Zhou, Fei Cheng, Xiaojun Zhang, and Kaizhu Huang. 2025. A parallel corpus of chinese-english legal judgments with argumentative structure annotations. *Data Intelligence*, 7(4):1291–1304.
- Zhangquan Zhou and Shijiao Tang. 2024. Investigating the logical capability of graph neural networks via the connection to c2. *Data Intelligence*, 6(3):720–748.

A Additional Methodological Details

A.1 Canonical Pair Classification and Ordered Edge Scoring

In the main pairwise causal classification task, each candidate event pair is represented under a fixed canonical order. The two possible causal directions are encoded by the CAUSE and EFFECT labels within the same three-way label space. Therefore, the main classifier does not enumerate (e_i, e_j) and (e_j, e_i) as two independent supervised instances.

During graph augmentation, the completion scorer may be applied to ordered event inputs to evaluate candidate directed edges. This ordered

scoring is used only to propose and weight intermediate graph edges before message passing. It does not introduce additional supervised training instances for the main classifier and is not used as a separate final decoder.

A.2 Scope of the Constrained Edge Completion Module

The Constrained Edge Completion Module (CECM) is designed for supervised directional edge proposal within causal graph refinement, rather than for unsupervised causal discovery over variables or causal mechanisms. It is trained with pairwise causal labels, but its role is structural: it calibrates retained classifier-derived edges and proposes a small number of reliable missing edges under structural constraints.

CECM contains a directional classifier and a direction-agnostic existence estimator. The directional classifier estimates the causal direction, while the existence estimator predicts whether a causal edge exists regardless of direction. Their outputs are combined into the existence-aware directional confidence η , as defined in Eq. 5. Therefore, the existence estimator is not used only as an auxiliary training signal; it also modulates edge activation and message-passing weights.

Specifically, for retained classifier-derived edges, η is fused with the main classifier confidence to produce a calibrated edge weight. For missing inter-sentence edges, CECM proposes only high-confidence candidate edges whose η scores satisfy the completion threshold, sentence-distance window, and per-source top- k constraints. In both cases, the module affects final prediction only through graph refinement and representation update, rather than through direct classifier ensembling or separate final decoding.

A.3 Reverse-Order Consistency for Proposed Edges

For a newly proposed edge $e_u \rightarrow e_v$, the forward ordered confidence η_{uv}^C measures whether the ordered input (e_u, e_v) supports the causal direction $e_u \rightarrow e_v$ with both directional and existence evidence. We additionally use η_{vu}^E as a reverse-order consistency signal. Under the reverse ordered input (e_v, e_u) , the EFFECT label also corresponds to the same causal direction $e_u \rightarrow e_v$. Therefore, a reliable proposed edge should receive consistent support from both input orders.

The consistency-based confidence is computed as:

$$a_{uv}^{comp} = \frac{1}{2} (\eta_{uv}^C + \eta_{vu}^E). \quad (22)$$

This confidence favors proposed edges whose directions are stable under both forward and reverse input orders and whose existence is also supported by the direction-agnostic edge-existence estimator. It is used only as the message-passing weight of the proposed edge in the refined graph. The edge activation decision itself is based on the forward existence-aware confidence η_{uv}^C together with the sentence-distance window and per-source top- k constraints. Thus, reverse-order consistency calibrates the confidence of proposed edges, but does not introduce additional supervised instances or serve as a separate final decoder.

A.4 Scope of the Conditional Diffusion Module

The Conditional Diffusion module does not generate textual content or graph structures from pure noise. Instead, it performs denoising over edge-level latent representations and provides a denoising-based support signal for retained candidate edges.

This design differs from using diffusion as a causal-direction classifier. Causal direction is determined by the main pairwise classifier for classifier-induced edges or by CECM for completed edges. CDM only adjusts the confidence of retained edges before graph reasoning. Therefore, CDM contributes to weak-edge reinforcement, while edge direction and edge activation remain controlled by the pairwise predictor and CECM.

A.5 Refined Graph as an Intermediate Reasoning Structure

The refined graph is not used as a separate final decoder. Instead, it serves as an intermediate reasoning structure that controls message passing among event representations. After each refinement round, the graph reasoner updates event representations, and the updated representations are reused by the pairwise classifier in the next round.

The final prediction is always produced by the pairwise classifier at the last executed refinement round. This design avoids directly treating intermediate graph edges as final labels, while still allowing structural refinement to influence prediction through representation learning.

Accordingly, the refinement modules influence final predictions only through intermediate graph construction and representation update, rather than through separate final decoding.

A.6 Clarification on Auxiliary Prediction and Diffusion Degeneracy

This section further clarifies the functional roles of CECM and CDM. The goal is to avoid two possible misunderstandings: first, that CECM is simply used as another supervised final classifier; and second, that CDM is merely a generic conditional autoencoder or an additional MLP regularizer.

Clarifying the role of CECM. Although CECM is supervised with pairwise causal labels, it is not used as an additional final classifier. Its outputs are consumed only during intermediate graph refinement. For retained classifier-induced edges, CECM calibrates edge confidence by combining directional confidence with direction-agnostic existence confidence. For missing inter-sentence edges, CECM proposes a small number of high-confidence directional candidates under the sentence-distance window and per-source top- k constraints. These proposed edges are used only as intermediate message-passing edges. Therefore, CECM affects final prediction through graph topology and representation updates, rather than through direct classifier ensembling or separate final decoding.

Clarifying the role of CDM. CDM is not designed to infer causal directions, activate new edges, or generate graph structures from noise. Causal directions are determined by the pairwise classifier for retained edges and by CECM for completed edges. CDM operates only on edge-level directional latent representations and converts the denoising result into a support gate for retained edges. Since the support gate is added to the calibrated edge confidence, CDM should be interpreted as providing positive confidence support rather than explicitly suppressing unreliable edges. Edges with more stable denoised latents receive stronger amplification, whereas unstable edges receive limited additional support.

Relation to autoencoder-style reconstruction. The denoising objective in CDM may appear related to conditional reconstruction because the denoiser is conditioned on the clean edge latent. We do not claim that the diffusion objective alone

discovers causal structure. Instead, its role is to provide a noise-conditioned consistency signal over edge-level representations before confidence-aware graph reasoning. This design differs from a generic autoencoder in its use of stochastic perturbation, timestep-conditioned denoising, relation-aware processing, and its restricted downstream use as an edge-confidence support gate rather than a reconstructed feature for direct classification.

Scope of the clarification. The above clarification does not rely on the assumption that diffusion denoising is mathematically incapable of degenerating into a simpler conditional mapping. Rather, it specifies how CDM is used in ICGR: CDM does not produce final labels, does not decide edge direction, and does not add new edges. Its only role is to provide an additional confidence-support signal for retained edges before message passing. We further examine this design choice through diagnostic variants in Appendix A.7, including a Noise-Augmented Gate variant that isolates stochastic perturbation from timestep-conditioned denoising.

A.7 Diagnostic Variants for CECM and CDM

To further examine whether the improvements of ICGR come from the proposed graph-refinement mechanisms rather than from additional prediction capacity or simple stochastic regularization, we conduct three diagnostic variants for CECM and CDM. These variants are designed to address three potential concerns: whether CDM behaves merely as an additional MLP-based regularizer, whether its gain can be explained by stochastic feature perturbation alone, and whether CECM behaves merely as another supervised classifier.

CDM $\rightarrow$ MLP Gate. In this variant, we replace the Conditional Diffusion Module with a parameter-matched MLP gate. The MLP gate takes the same edge-level directional representation as input and directly predicts a confidence-support score for retained edges. Unlike CDM, this variant does not perform stochastic noising, timestep-conditioned denoising, or relation-aware reverse denoising. The resulting support score is added to the calibrated edge confidence in the same way as the diffusion-based gate. This comparison tests whether the gain of CDM can be explained by an additional nonlinear gating function alone.

CDM $\rightarrow$ Noise-Augmented Gate. In this variant, we keep the stochastic perturbation step but

remove the timestep-conditioned reverse denoising process and the diffusion loss. Specifically, the clean edge latent representation is perturbed with Gaussian noise, and the confidence-support gate is directly predicted from the clean and noisy latents. The resulting gate is added to the calibrated edge confidence in the same way as CDM. This comparison tests whether the gain of CDM can be explained by noise augmentation alone, without diffusion-style denoising.

CECM as Ensemble. In this variant, we use CECM as an auxiliary supervised predictor, but disable its graph-refinement pathway. Specifically, CECM predictions are not used to calibrate retained edge weights, propose missing edges, or modify the intermediate message-passing graph. Instead, the CECM directional scores are ensembled with the main pairwise classifier only at the prediction level. This comparison tests whether the gain of CECM mainly comes from its role as another supervised classifier or from its use in refining the intermediate causal graph.

Results. Table 4 reports the diagnostic results. Compared with removing CDM, the MLP gate variant improves performance, indicating that direct edge-level confidence support is useful. The Noise-Augmented Gate variant further improves over the MLP gate, suggesting that stochastic perturbation can provide additional robustness for retained edge representations. However, it still remains below the full CDM, which indicates that the gain of CDM cannot be fully explained by either additional nonlinear gating capacity or noise augmentation alone. The timestep-conditioned denoising process provides a stronger stability signal for weak retained edges. Similarly, compared with removing CECM, the ensemble variant uses the same supervised CECM scores, but does not allow them to affect graph topology or message passing. Therefore, the relative performance of these variants helps distinguish confidence-conditioned graph-refinement effects from auxiliary prediction or simple regularization effects.

Discussion. These diagnostic variants are intended to clarify the functional roles of CECM and CDM. The CDM $\rightarrow$ MLP Gate variant improves over w/o CDM but remains below the full model, suggesting that direct edge-level gating is helpful but cannot fully replace diffusion-based confidence support. The CDM $\rightarrow$ Noise-Augmented Gate vari-

Model	Overall F1	
	Direction	Existence
ICGR (Full Model)	50.3	65.0
w/o CDM	45.7	58.8
CDM $\rightarrow$ MLP Gate	47.7	61.4
CDM $\rightarrow$ Noise-Augmented Gate	48.6	62.9
w/o CECM	46.8	60.5
CECM as Ensemble	48.6	61.9

Table 4: Diagnostic variants for CECM and CDM on EventStoryLine. The CDM $\rightarrow$ MLP Gate variant replaces diffusion-based denoising with a parameter-matched direct gating function. The CDM $\rightarrow$ Noise-Augmented Gate variant perturbs the clean edge latent representation with Gaussian noise and directly predicts a confidence-support gate without timestep-conditioned reverse denoising or diffusion loss. The CECM as Ensemble variant uses CECM as an auxiliary predictor without allowing it to refine the intermediate graph topology or message-passing weights.

ant further improves over the MLP gate, showing that stochastic perturbation itself contributes to robustness (Gao et al., 2025). Nevertheless, its performance is still lower than that of the full CDM, indicating that timestep-conditioned denoising provides additional benefits beyond noise augmentation alone. The CECM as Ensemble variant improves over w/o CECM but remains below the full model, suggesting that CECM is more effective when its predictions are used to refine the intermediate graph structure rather than being directly ensembled as final classification scores. These results support the view that the gains of ICGR come from confidence-conditioned graph refinement instead of merely adding auxiliary predictors or simple stochastic regularization.

B Implementation Details

The proposed model is implemented in PyTorch with bert-base-uncased as the backbone encoder. For each event mention, we represent its trigger by averaging the contextualized token embeddings within the trigger span.

Optimization and Architecture. We optimize the model with AdamW and use differential fine-tuning for the pre-trained encoder and task-specific modules. The learning rates are 2×10^{-5} for BERT and 1×10^{-4} for non-BERT modules. The model is trained for 30 epochs with a batch size of 1. To alleviate class imbalance, we use focal loss with focusing parameter $\gamma = 2.0$. For graph reasoning, we use a 2-layer CGE network with 4 attention heads.

Component	Parameter	Value
Sparse graph construction	ω	0.6
CECM	τ_{disc} on η, W , top- k	0.9, 3, 2
Completion calibration	$\alpha_{disc}, \lambda_{conf}$	0.25, 0.1
CDM	T, K_{intra}, K_{inter}	100, 3, 12
Diffusion support	$\lambda_{intra}, \lambda_{inter}$	0.05, 0.08
Graph reasoning	μ, τ	1.0, 1.0
Training objective	$\lambda_{disc}, \lambda_{diff}$	0.15, 0.05
Adaptive refinement	N_{min}, N_{max}, δ	2, 9, 2

Table 5: Values of equation-level parameters used in the main experiments. We report parameters that appear in the model equations or graph-construction rules, while omitting low-level implementation constants that are not part of the method definition.

All reported results use the same optimization and architecture settings unless otherwise specified.

Equation-Level Settings. To maintain reproducibility without exposing unnecessary low-level implementation details, we report the values of parameters that appear in the model equations or graph-construction rules. All structural hyperparameters were selected only on the development set and then fixed for all reported test evaluations. These equation-level settings are kept fixed across the main experiments unless otherwise specified.

Table 5 summarizes the equation-level parameters used in the main experiments. For CECM, τ_{disc} is applied to the existence-aware directional confidence η in Eq. 5; no additional threshold is introduced.

B.1 Additional Details of Diffusion Inference

During inference and iterative refinement, the Conditional Diffusion module starts from a lightly noised version of the current edge latent. We use a standard linear noise schedule with $T = 100$ diffusion timesteps. For a relation type $r \in \{intra, inter\}$, we set the starting denoising timestep to K_r and initialize:

$$\mathbf{z}_{ij}^{K_r, (t)} = \sqrt{\bar{\alpha}_{K_r}} \mathbf{z}_{ij}^{0, (t)} + \sqrt{1 - \bar{\alpha}_{K_r}} \epsilon, \quad (23)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

For each reverse step $k = K_r, \dots, 1$, the relation-specific denoiser predicts the injected noise as:

$$\hat{\epsilon}_k = \epsilon_\theta^r \left(\mathbf{z}_{ij}^{k, (t)}, \mathbf{z}_{ij}^{0, (t)}, k \right). \quad (24)$$

We then estimate the clean latent by:

$$\hat{\mathbf{z}}_{ij}^{0, (t)} = \frac{\mathbf{z}_{ij}^{k, (t)} - \sqrt{1 - \bar{\alpha}_k} \hat{\epsilon}_k}{\sqrt{\bar{\alpha}_k}}. \quad (25)$$

The next latent is computed deterministically as:

$$\mathbf{z}_{ij}^{k-1, (t)} = \sqrt{\bar{\alpha}_{k-1}} \hat{\mathbf{z}}_{ij}^{0, (t)} + \sqrt{1 - \bar{\alpha}_{k-1}} \hat{\epsilon}_k. \quad (26)$$

After the final reverse step, the resulting latent is denoted as $\hat{\mathbf{z}}_{ij}^{0, (t)}$ and is used only to compute the support gate in Eq. 14. It is not used to generate graph edges or predict causal direction directly.

C Baselines

For PLM-based methods, we employ **BERT** (Devlin et al., 2018) and **RoBERTa** (Liu et al., 2019). Regarding structure-aware methods, we include **ERGO** (Chen et al., 2022), which models event interactions via a relational graph Transformer, and **SENDIR** (Yuan et al., 2023), which facilitates inter-sentence reasoning by constructing causal chains from sparse representations. We also compare with **iLIF** (Liu et al., 2024), an iterative framework that updates graph structures dynamically, and the recent **NCHM** (Xiang et al., 2025), which captures high-order dependencies using hypergraph convolution. To assess generative paradigms, we incorporate **DiffusECI** (Man et al., 2024), a diffusion-based method focusing on feature-level denoising.

For the main comparison tables, we follow the standard reporting practice in document-level ECI studies. In particular, the EventStoryLine comparisons follow the common topic-level evaluation protocol (Phu and Nguyen, 2021; Gao et al., 2019), and the MAVEN-ERE comparisons follow the pre-processing and evaluation setting adopted in prior work Tao et al. (2023); Liu et al. (2024).

Finally, we benchmark against Large Language Models (LLMs), including previously reported LLM results and our additional evaluations of Llama-3.1-8B⁴ and Qwen3-Max⁵. We treat the LLM results as reference comparisons for the generative prompting paradigm, rather than as fully controlled comparisons against supervised SOTA systems. Detailed LLM settings are provided in Appendix G.

D Dataset Details

We evaluate the proposed framework on two representative benchmarks: EventStoryLine (v0.9) and MAVEN-ERE.

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B>

⁵<https://qwen.ai/blog?id=qwen3-max>

D.1 EventStoryLine (ESL)

The **EventStoryLine (v0.9)** dataset (Caselli and Vossen, 2017) contains 258 documents across 22 topics, comprising 5,334 event mentions. It includes 1,770 intra-sentence and 3,855 inter-sentence causal pairs, making it suitable for evaluating long-distance dependencies. For data preprocessing, we follow Liu et al. (2024), mapping ‘PRECONDITION’ relations to the CAUSE label and ‘FALLING_ACTION’ relations to the EFFECT label. For evaluation, the last two topics are reserved as the development set and the remaining 20 topics are used for 5-fold cross-validation. We report the average Precision (P), Recall (R), and F1 scores over the five folds.

D.2 MAVEN-ERE

The **MAVEN-ERE** dataset (Wang et al., 2022) is a large-scale benchmark containing 4,480 documents, 103,193 event mentions, and 57,992 causal event pairs. In the original annotations, causal event pairs are labeled as ‘CAUSE’ or ‘PRECONDITION’, both of which indicate causal relations. For the direction identification setting, we follow the preprocessing and evaluation protocol of Liu et al. (2024) so that our results are directly comparable with prior work under the same benchmark configuration. Following this shared protocol, the original causal annotations are converted into a direction-identification benchmark.

Since MAVEN-ERE does not provide an official test set, following Tao et al. (2023); Liu et al. (2024), we use the original development set as the test set and further sample 10% of the original training data as a new development set.

For each experiment, we report the average Precision (P), Recall (R), and F1 scores over five random seeds. This setup allows us to maintain direct comparability with previously published results while providing a stable evaluation of model performance.

D.3 Ablation Study on MAVEN-ERE

To further analyze the contribution of each component, we conduct an additional ablation study on MAVEN-ERE for causality direction identification. To ensure fair comparison with prior work, we follow the same preprocessing protocol as Liu et al. (2024). Specifically, we evaluate the impact of removing iterative refinement, the Constrained Edge Completion module, and the Conditional Diffusion

Model	Intra			Inter			Overall		
	P	R	F1	P	R	F1	P	R	F1
ICGR (Ours)	73.3	52.5	61.2	65.9	50.2	57.0	65.8	51.3	57.7
w/o Iteration	53.6	48.6	50.9	48.4	46.0	47.2	48.8	47.3	48.4
w/o CECM	60.1	50.1	54.6	57.7	49.5	53.2	57.9	49.8	53.5
w/o CDM	60.1	50.3	54.8	47.5	48.8	48.1	48.4	49.1	48.8

Table 6: Ablation results on MAVEN-ERE for causality direction identification. We report precision (P), recall (R), and F1 (%).

module, respectively.

Table 6 shows that all three components remain beneficial on MAVEN-ERE. Removing iterative refinement produces the largest overall degradation, again confirming the importance of multi-round structural refinement for document-level causal reasoning. Removing the Conditional Diffusion module also leads to a substantial drop, indicating that diffusion-based denoising and weak-edge support remain important for stabilizing directional predictions under noisy document context. Finally, removing the Constrained Edge Completion module yields a consistent but smaller degradation, showing that constrained directional causal edge recovery is still beneficial even when strong pairwise and graph-based signals are already available.

Taken together, the ablation results on EventStoryLine and MAVEN-ERE are highly consistent: the full model performs best because the Conditional Diffusion module, the Constrained Edge Completion module, and iterative refinement address complementary aspects of document-level event causality identification.

E Additional Exploration Study

We further analyze how the causal graph evolves throughout the iterative refinement process of ICGR. Since our framework combines confidence calibration for weak retained edges with constrained directional causal edge completion, the intermediate graph states provide a direct view of whether sparse document-level causal structure becomes more complete and more coherent over successive refinement rounds.

Figure 7 compares the initial graph and the final refined graph on the EventStoryLine test set. For both causality direction identification and causality existence identification, the final graph contains a higher proportion of identified causal relations and a lower proportion of un-identified causal relations than the initial graph. This improvement is more evident for inter-sentence relations, which

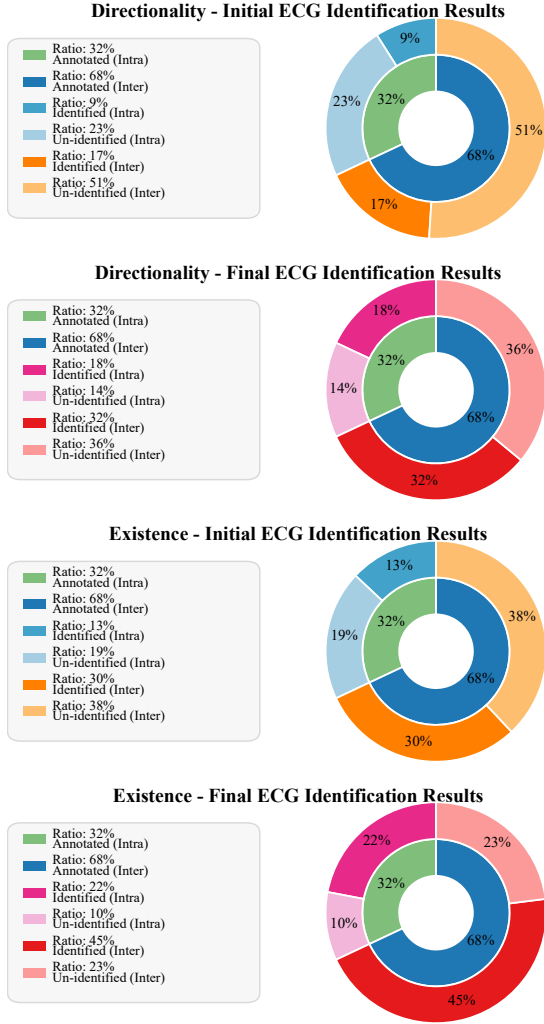


Figure 7: Comparison between the initial graph and the final refined graph on the EventStoryLine test set. We report the proportions of identified and un-identified annotated causal relations for intra-sentence and inter-sentence cases under both direction identification and existence identification settings. The final refined graph substantially reduces un-identified relations, especially for inter-sentence causality.

are typically associated with weaker local cues and longer reasoning paths. The comparison indicates that iterative refinement helps recover part of the causal structure that is difficult to preserve under a single conservative decoding step, especially in the cross-sentence setting.

Figure 8 provides a concrete example of this process.

At initialization, the graph misses several gold long-distance relations and also contains a few incorrect edges. As refinement proceeds, weak but structurally useful links receive stronger support through confidence-aware reasoning, allowing downstream events to reconnect to a more complete causal chain. Meanwhile, some early incor-

Model	Overall F1 (%)
iLIF	46.69 $\pm$ 2.65
ICGR (Ours)	50.30 $\pm$ 1.34

Table 7: Fold-level statistical comparison on EventStoryLine. Results are reported as mean $\pm$ standard deviation over the 5 folds. Paired t -test: $p = 0.0035$.

rect directions are corrected after later rounds provide stronger structural and document-level context. The final graph is therefore substantially closer to the gold structure than the initial graph, qualitatively supporting the effectiveness of ICGR in improving sparse document-level causal topology.

F Statistical Significance Analysis

To further examine whether the improvement of ICGR is robust under a controlled setting, we compare ICGR with the strongest reproduced baseline, **iLIF**. Following the standard EventStoryLine 5-fold topic-level evaluation protocol used in prior work, we conduct paired significance testing over the same five folds. For this controlled comparison, iLIF is reproduced using the officially released code and the official environment specification, and ICGR is evaluated under the same unified environment configuration and evaluation pipeline to ensure a fair comparison. Table 7 reports the fold-level overall F1 scores in terms of mean and standard deviation. ICGR achieves a higher average F1 than iLIF (**50.30 $\pm$ 1.34** vs. 46.69 $\pm$ 2.65), while also showing lower variance across folds, indicating more stable performance.

We further conduct paired significance testing over the 5 folds. Under this controlled setting, a paired t -test yields $p = 0.0035$, indicating that the fold-level improvement of ICGR over iLIF is statistically significant under this test.

G Details of LLM Baselines

To benchmark the generative paradigm, we incorporate a diverse set of Large Language Models (LLMs). We reference the results of *text-davinci-003*, *GPT-3.5-turbo*, and *GPT-4* from prior work (Gao et al., 2023). In addition, to reflect the performance of more recent LLMs, we conduct independent evaluations on *LLaMA-3.1-8B* and *Qwen3-Max*.

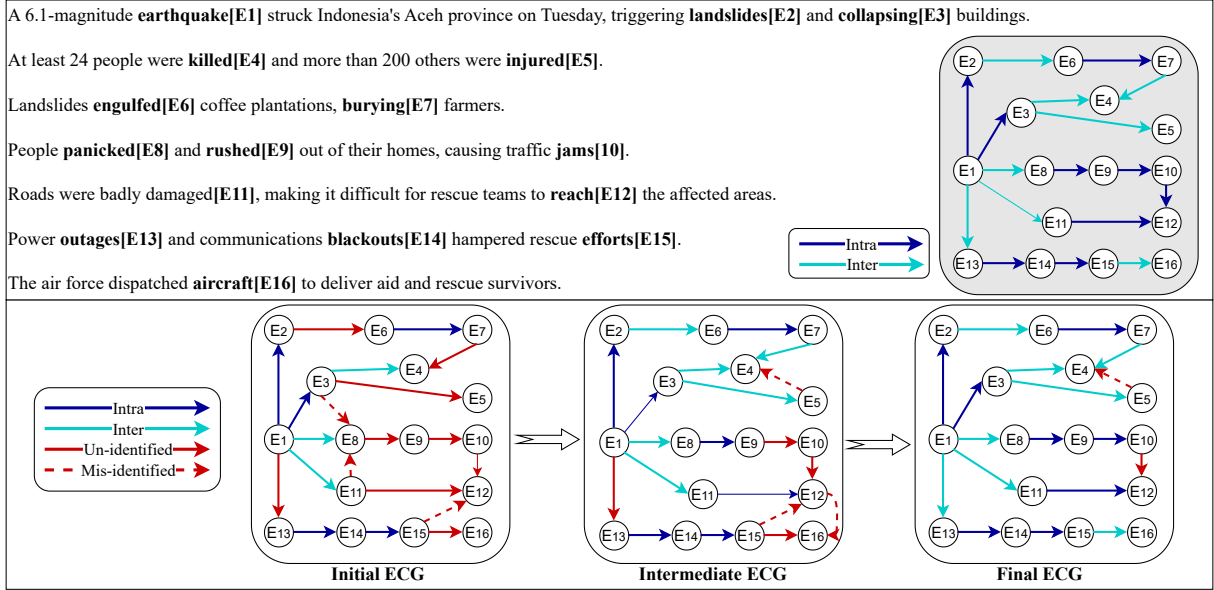


Figure 8: A representative case of iterative causal graph refinement by ICGR. Compared with the annotated causal graph, the initial graph contains several missing and spurious causal links. As refinement proceeds, some weak long-distance relations are gradually recovered and some incorrect edges are corrected. The final refined graph is much closer to the gold graph, demonstrating the ability of ICGR to restore a more coherent document-level causal structure.

G.1 Experimental Setting

We evaluate LLM baselines under a prompt-based event causality identification protocol following the task formulation and prompt design principles of Gao et al. (2023). Rather than treating prompt engineering as an exhaustive search for the optimal prompt, our goal is to assess whether commonly used instruction, in-context learning (ICL), and chain-of-thought (CoT) prompting paradigms can provide reliable fine-grained causal discrimination under the same event-pair evaluation setting (He et al., 2025).

For each candidate event pair, the prompt presents the document context, explicitly marks the two target events, defines the causality identification question, and requires the model to produce a structured answer. For local LLM evaluation, we use *LLaMA-3.1-8B-Instruct* and evaluate both zero-shot CoT and few-shot CoT settings. In the zero-shot CoT setting, the model is asked to reason step by step and then output “Answer: Yes” or “Answer: No”. In the few-shot CoT setting, we use 12 demonstrations, including 4 positive and 8 negative examples, following the reasoning template of Gao et al. (2023): event extraction, semantic association analysis, optional commonsense reasoning, causal direction or non-causality verification, and final structured output.

We parse model outputs using deterministic an-

swer extraction rules over the text following “Answer:” and compute precision, recall, F1, and accuracy for intra-sentence, inter-sentence, and overall settings.

G.2 Evaluation Scope of LLM Baselines

The LLM results are intended as reference comparisons for the generative prompting paradigm, rather than fully controlled comparisons against supervised SOTA systems. In Section 4.2, the evaluation scope differs across LLM baselines. For LLaMA-3.1-8B, we report results under intra-sentence, inter-sentence, and overall settings on EventStoryLine. For Qwen3-Max, we report only intra-sentence results, while leaving inter-sentence and overall results unevaluated due to computational cost. For some earlier LLM baselines, we directly report previously published results when available.

Task Reliability of Current LLMs. The LLM results should be interpreted as diagnostic evidence about prompt-based LLM behavior on fine-grained ECI, rather than as a definitive upper bound on all possible LLM-based methods. Prior work has shown that ChatGPT-style models often obtain high recall but low precision on ECI, suggesting a tendency to over-predict causal relations; it also reports that ICL and CoT prompting may further exacerbate this imbalance, and that performance

Model	P (%)	R (%)	F1 (%)
GPT-4 (Gao et al., 2023)	27.2	94.7	42.2
Qwen3-Max (zero-shot)	7.2	87.6	13.3
Qwen3-Max (few-shot)	38.5	76.5	51.2

Table 8: Supplementary LLM results on EventStoryLine under the intra-sentence causality existence identification setting. Since inter-sentence and overall document-level results are not available for these models, they are reported as partial reference comparisons.

is sensitive to the wording and format of causal prompts (Gao et al., 2023).

Our local LLM experiments follow this evaluation protocol and exhibit the same failure pattern. In particular, under zero-shot CoT on EventStoryLine, *LLaMA-3.1-8B-Instruct* achieves very high recall but extremely low precision, especially in the inter-sentence setting. This indicates that the model frequently recognizes potential causal signals but struggles to reject non-causal event pairs. Since document-level ECI involves longer contexts, denser event interactions, and more implicit causal evidence, such over-prediction behavior is especially problematic for fine-grained causal discrimination.

Therefore, we use LLM baselines as reference comparisons for the generative prompting paradigm. The results do not imply that future LLMs or task-specific LLM training cannot improve DECI performance; instead, they show that standard prompt-based LLM inference, even with CoT and few-shot demonstrations, is not yet a reliable substitute for supervised document-level causal structure modeling.

G.3 Supplementary Intra-sentence LLM Results

In addition to the LLM comparisons discussed above, we report supplementary results for several LLM baselines under the intra-sentence causality existence identification setting on EventStoryLine. These results are included in the appendix because the corresponding inter-sentence and overall document-level results are not available for all models. Therefore, they should be interpreted as partial reference comparisons rather than fully controlled document-level comparisons.

As shown in Table 8, LLMs generally obtain high recall but relatively low precision in this setting, indicating that they tend to over-predict causal relations. Few-shot prompting improves Qwen3-

Max compared with the zero-shot setting, but its F1 score remains lower than that of ICGR under the same intra-sentence existence identification setting. These results suggest that although LLMs can capture broad causal associations, they still struggle with fine-grained causal discrimination.

G.4 Prompt Templates for LLM Baselines

To improve reproducibility and clarify the scope of our LLM evaluation, we report the prompt templates used for the local LLM baselines. These prompts follow the event-pair causality identification formulation of Gao et al. (2023), where the model is given the document context and two target events, and is required to predict whether a causal relation exists between them.

Zero-shot CoT prompt.

Input: {doc_content}
Event 1: "{event1}"
Event 2: "{event2}"
Question: is there a causal relationship between "{event1}" and "{event2}"?
Let’s think step by step. Then state your answer as “Answer: Yes” or “Answer: No”.
Answer:

Few-shot CoT prompt. For the few-shot CoT setting, we prepend 12 demonstrations to the test instance, including 4 positive and 8 negative examples. Each demonstration follows a six-step reasoning format: extracting the two event descriptions, analyzing their semantic association, optionally using commonsense knowledge, verifying causal direction or non-causality, and producing a final structured answer.

{icl_examples}
Input: {doc_content}
Event 1: "{event1}"
Event 2: "{event2}"
Question: is there a causal relationship between "{event1}" and "{event2}"?
Reasoning Process:

We parse the model output by extracting the first valid “Yes” or “No” after the keyword “Answer:” for causality existence prediction. Outputs that cannot be parsed into either label are treated as invalid and excluded from metric computation.

Method	Time / Epoch	GPU Memory / Batch
SENDIR (Yuan et al., 2023)	2362 s	24 GB
ERGO (Chen et al., 2022)	131 s	24 GB
iLIF (Liu et al., 2024)	823 s	16 GB
ICGR (Ours)	134 s	18 GB

Table 9: Training efficiency comparison with representative document-level ECI baselines. "Time / Epoch" denotes the average wall-clock training time per epoch, and "GPU Memory / Batch" denotes the peak GPU memory consumption under the corresponding batch setting.

H Evaluation Settings and Label Configurations

To comprehensively assess the model’s capabilities in both fine-grained role differentiation and fundamental link completion, we conduct experiments under two evaluation settings.

Causality Direction Identification This setting evaluates the model’s ability to distinguish causal directionality. Each directed event pair is assigned one of three labels: NONE, CAUSE, or EFFECT.

Causality Existence Identification This setting focuses on determining whether a causal link exists between an event pair, regardless of its direction. For this setting, CAUSE and EFFECT are collapsed into a single positive class, while NONE is treated as the negative class. A prediction is counted as correct if it correctly determines whether the event pair is causal or non-causal, without requiring the predicted causal direction to match the gold direction. For MAVEN-ERE, following prior work, we obtain the existence identification results by merging the event pairs predicted as CAUSE and EFFECT in the direction setting into the CAUSAL class.

I Computational Efficiency and Resource Usage

We further compare the computational efficiency of representative document-level ECI methods. As shown in Table 9, SENDIR (Yuan et al., 2023), ERGO (Chen et al., 2022), and iLIF (Liu et al., 2024) exhibit substantially different training costs in terms of epoch time and GPU memory usage. This comparison provides a practical reference for understanding the efficiency–effectiveness trade-off of structure-aware and iterative ECI models.

The results for ICGR are measured on a single NVIDIA GeForce RTX 3090 GPU with 24GB memory, while the baseline efficiency numbers are reported following their available implementations or published settings when applicable.

As shown in Table 9, although ICGR adopts iterative graph refinement together with relation-aware denoising, the overall training cost remains manageable in practice. Under the default setting of 30 training epochs, one complete training run takes approximately 1.0–1.2 hours, and the full 5-fold cross-validation on EventStoryLine can be completed in about 5.7 hours. In addition, the peak GPU memory usage is around 18GB, indicating that the proposed framework does not exhaust the memory budget of the hardware used in our experiments.

Overall, these results suggest that the additional cost introduced by iterative refinement is acceptable for document-level ECI, while still leaving room for future improvement in more resource-constrained or latency-sensitive scenarios.