# Efficient Document-level Event Relation Extraction

**Ruochen Li, Zimu Wang, Xinya Du**
University of Texas at Dallas
{ruochen.li, zimu.wang, xinya.du}@utdallas.edu

## Abstract

Event Relation Extraction (ERE) predicts temporal and causal relationships between events, playing a crucial role in constructing comprehensive event knowledge graphs. However, existing approaches based on pairwise comparisons often suffer from computational inefficiency, particularly at the document level, due to the quadratic operations required. Additionally, the predominance of unrelated events also leads to largely skewed data distributions. In this paper, we propose an innovative two-stage framework to tackle the challenges, consisting of a retriever to identify the related event pairs and a cross-encoder to classify the relationships between the retrieved pairs. Evaluations across representative benchmarks demonstrate our approach achieves better efficiency and significantly better performance. We also investigate leveraging event coreference chains for ERE and demonstrate their effectiveness.

## 1 Introduction

Event Relation Extraction (ERE) aims at identifying relationships between events, especially temporal and causal connections. As illustrated in Figure 1, given the original text and three event mentions of interest, an ERE model should detect and classify the temporal (e.g., *overlaps* and *before*) and causal (e.g., *cause*) relationships between them. ERE plays a pivotal role in the construction of event knowledge graphs (EKGs, Ma et al., 2022) and supports a variety of tasks, such as future event prediction (Lin et al., 2022), machine reading comprehension (Zhu et al., 2023), and multi-hop reasoning (Li et al., 2024).

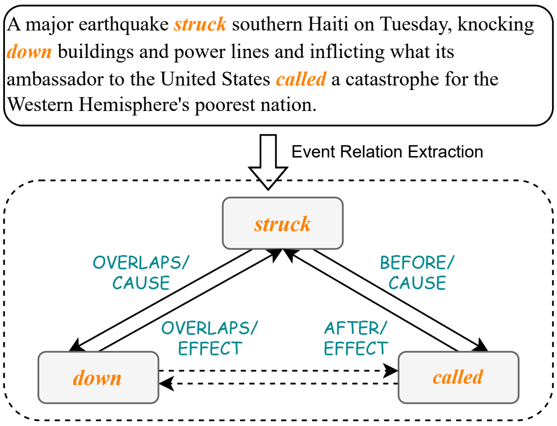
A major earthquake *struck* southern Haiti on Tuesday, knocking *down* buildings and power lines and inflicting what its ambassador to the United States *called* a catastrophe for the Western Hemisphere's poorest nation.

Figure 1: An example of ERE task with temporal and causal relations. The dashed lines indicate there are no event relations between the event mentions.

ERE is challenging due to the event relation variety and the required comprehension (Liu et al., 2020b). For document-level ERE (DERE), the challenge intensifies, needing event disambiguation and connection across expansive narrative structures. Previous research has mainly focused on enriching event semantics (Wen and Ji, 2021; Tran Phu and Nguyen, 2021), or exploiting large language models (LLMs) (Peng et al., 2023a). Nevertheless, current research faces a unique challenge in inefficient learning and inference because the determination of relationships requires pairwise classification after iterating through *all event pairs* (Hu et al., 2023; Wang et al., 2024), which inherently exhibits quadratic time complexity. Additional training challenges arise due to the largely skewed data distribution, where most event pairs have no relation, becoming particularly critical for DERE with a broader scope of events and lengthy sources (Gao et al., 2023). However, this aspect has been overlooked in existing studies, and we are the first to investigate the efficiency issue in DERE with crucial yet unexplored temporal and causal relations.

In this paper, we introduce a novel pruning-based two-stage paradigm for DERE (Figure 2). In the first stage of the framework, we employ a retriever model to efficiently sift through event mentions in latent embedding spaces and identify the related event pairs. Afterwards, a cross-encoder is fine-tuned for event relation prediction on the narrowed set of candidate pairs. This approach effectively prunes the candidate event pairs to tackle

*Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 92–99
May 4, 2025 ©2025 Association for Computational Linguistics

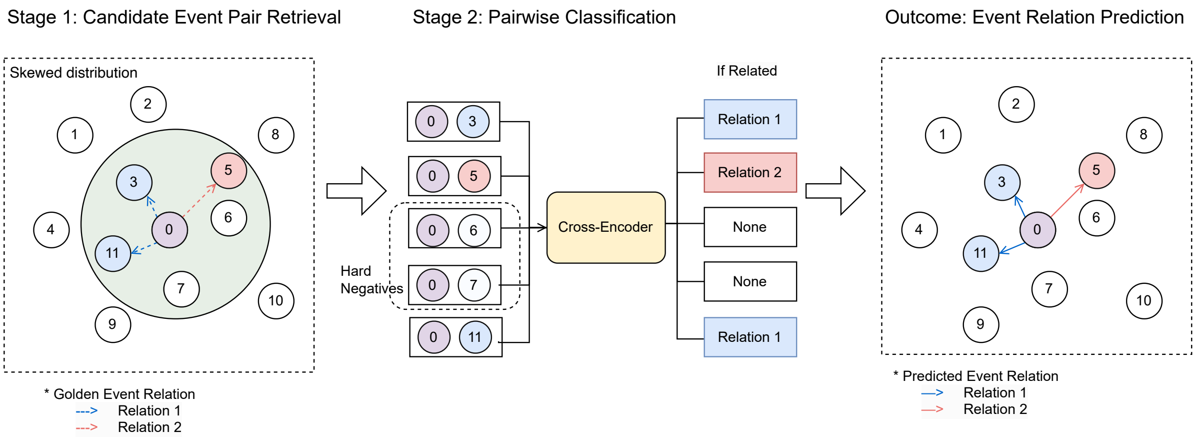

Figure 2: Overall architecture of the proposed pruning-based two-stage ERE framework. Stage 1 retrieves candidate event pairs, and then Stage 2 conducts the costly fine-grained event relation predictions on the retrieved pairs.

inefficiency and deals with the skewed distribution to enhance performance. Experimental results on Event StoryLine Corpus (ESC, Caselli and Vossen, 2017), Richer Event Description (RED, O'Gorman et al., 2016), and MAVEN-ERE (Wang et al., 2022) demonstrate significantly better performance and efficiency compared to representative baselines.

In summary, the key contributions of this paper are as follows: (1) We design a novel two-stage framework for DERE by pruning candidate event pairs to reduce computational complexity and mitigate the skewed distribution issue. (2) We conduct rigorous evaluations and ablation studies on DERE datasets with various retrievers and cross-encoders. (3) We conduct a comprehensive analysis, including time complexity, the effectiveness of encoding strategies and coreference chains, and the effect of retrieved candidate-pair count on performances.

## 2 Related Work

Recent progress of ERE has been made based on pre-trained language models (PLMs), utilizing semantic structures (Tran Phu and Nguyen, 2021; Hu et al., 2023), temporal clues (Wen and Ji, 2021), and external knowledge (Liu et al., 2020a; Cao et al., 2021) to enrich the event representations. Some other works leverage the high-order transitivity (Chen et al., 2022, 2023) and multi-task learning (Ning et al., 2018; Wang et al., 2022) to model the dependencies between different relation types. Some researchers further investigate the use of LLMs in ERE (Gao et al., 2023; Peng et al., 2023a,b; Wang et al., 2024). However, they often achieve this at the expense of computational efficiency when performing pairwise classifications, especially in document-level datasets (O'Gorman et al., 2016; Wang et al., 2022). While some recent work focus on improving the efficiency of entity coreference resolution (Lee et al., 2018; Held et al., 2021), they cannot be generalized to DERE because of the requirement for deeper semantic analysis and the existence of more specific relation types. In this paper, we inherit the ideas of pruning but design a more effective framework with a retriever model and a cross-encoder model.

## 3 Methodology

We formulate our DERE task as a multi-class classification problem. Formally, given a document $D$ that contains multiple sentences and two event mentions $e_h$ and $e_t$ of interest, our goal is to predict the potential temporal (e.g., *before*) and causal (e.g., *cause*) relationships between them. Following the framework shown in Figure 2, we introduce the implementation of the retriever and cross-encoder models for training and inference in detail.

### 3.1 Candidate Event Pair Retrieval

The initial stage utilizes a retriever model (i.e., *bi-encoder*[1]) to efficiently represents event mentions in a latent embedding space to identify the event pairs likely to have a relation to improve efficiency and alleviate the skewed distribution problem. Formally, for two events $e_h$ and $e_t$, the wrapped mentions are defined as the sentences containing them $(s_h, s_t)$ with events wrapped by markers `<m>` and `</m>` for enhanced emphasis. With bi-encoder denoted as $Enc(\cdot)$, the representation of the events, $\mathbf{r}_h$ and $\mathbf{r}_t$, are encoded as:

$$\mathbf{r}_h = Enc(s_h) = Enc(\texttt{<s>}\ldots\texttt{<m>}e_h\texttt{</m>}\ldots\texttt{</s>}), \quad (1)$$

[1] Bi-encoders are a broad class of models that map the input and candidate responses separately into a common feature space where their similarity is measured (Huang et al., 2021).

$$\mathbf{r}_t = Enc(s_t) = Enc(\texttt{<s>}\ldots\texttt{<m>}e_t\texttt{</m>}\ldots\texttt{</s>}). \quad (2)$$

Afterwards, we select the top 5 event mentions most likely to form a relationship with each event. For fine-tuning, the task is regarded as binary classification over $\mathbf{r}_h \cdot \mathbf{r}_t$ with cross-entropy loss. This process is intended to direct the model's attention to the most significant elements of the event, thereby improving its ability to discern relevant event pairs.

### 3.2 Pairwise Classification

In the second stage, we conduct pairwise classification on the pruned candidate set with a cross-encoder, for which we employ both discriminative and generative models.

**Discriminative Models.** Given an input document $D$, we first obtain the hidden vectors in the last transformer layer. Then, for event mentions $e_h$ and $e_t$, we compute the representations $\mathbf{r}_h$ and $\mathbf{r}_t$ by averaging the representation vectors of respect tokens. Finally, we form an overall representation vector $\mathbf{r}_{h\to t}$ by concatenating the two representations: $\mathbf{r}_{h\to t} = [\mathbf{r}_h; \mathbf{r}_t]$, and then feed it to a feed-forward neural network for relation classification.

**Generative Models.** The generative models utilize a Seq2Seq approach. An example is as follows:

> *Classify: Mention 1: The murder* <m> ***trial*** </m> *of a suspended female [...]* <sep> *Mention 2: The murder trial [...]* <m> ***shooting*** </m> *three co-workers [...]*

The design of the instruction starts from the word "*Classify:*". Then, we add the two sentences containing the events, separated by a special symbol <sep>, and we wrap the mentions with the markers <m> and </m>. <s> and </s> denote the sentence boundary. The output of the model is the specific event relationship between them.

All positive event pairs and *hard negatives*, i.e., the negatives retrieved in Stage 1, are used for training. We adopt this strategy because (1) the cross-encoder makes predictions on the outputs of the retriever, which are essential for its learning process; and (2) the retrieved event pairs are identified as related ones, making them ideal candidates for more expensive cross-comparison.

### 3.3 Inference

The inference process is distinct from training with additional strategies. Utilizing the trained retriever model, we retrieve a set of $k$ events most likely to form a relationship with the given mention, regardless of whether it is fine-tuned, and then we use the cross-encoder to determine the specific relationship between them. Our approach prunes aggressively to improve efficiency. Stage 1 scales linearly, and retrieved event pairs are sent to the cross-encoder (Stage 2). In this case, the time of quadratic operation can be decreased significantly, and both the prediction and efficiency can be improved.

## 4 Experiments

### 4.1 Datasets and Experimental Setup

We conduct experiments on three well-established datasets: Event StoryLine Corpus (ESC, Caselli and Vossen, 2017), Richer Event Description (RED, O'Gorman et al., 2016), and MAVEN-ERE (Wang et al., 2022). For MAVEN-ERE, we follow previous work (Gao et al., 2023; Chen et al., 2024) to sample a subset. We report the precision (P), recall (R), and micro F1-score (F1) under (1) event pairs with relations following previous work, and (2) all event pairs, as it is closer to EKG construction.

We employ diverse retrievers (RoBERTa-Large, Liu et al., 2019) and S-BERT (Reimers and Gurevych, 2019)) and classifiers (RoBERTa-Large and T5-Large (Raffel et al., 2020)). The baseline models we compared are in Appendix A.

### 4.2 Experimental Results

Experimental results are depicted in Tables 1 and 2, from which we have the following observations:

Firstly, the introduce of retriever significantly enhances cross-encoder performance, with BERT and RoBERTa outperforming more complex models like LIP and RichGCN. GPT-3.5 does not outperform PLM-based approaches due to its zero-shot generative nature. Additionally, our retriever also outperforms random sampling as it is more closely aligned with the inference process. The hard negatives identified by the retriever are also more similar to positive ones, which are hard to differentiate.

Secondly, while all metrics increase simultaneously, the recall values increase more, particularly on ESC (increases two times) containing large negative samples. Simultaneously, the performance on non-negatives has significant increase; thus, the skewed distribution problem can be alleviated. The T5 model with the S-BERT retriever achieves the best performance on all datasets, demonstrating their superior capability in event relation classification and candidate event pair identification.

| Retriever | Classifier | ESC |  |  | RED |  |  |
|---|---|---|---|---|---|---|---|
|  |  | P | R | F1 | P | R | F1 |
| Random | RoBERTa | 78.4 | 85.2 | 81.7 | 81.2 | 86.2 | 83.6 |
|  | T5 | 86.1 | 85.7 | 85.9 | 86.8 | 89.9 | 88.3 |
| RoBERTa | RoBERTa | 77.2 | 86.9 | 81.8 | 82.7 | 87.5 | 85.0 |
| RoBERTa (Fine-tuned) | RoBERTa | 77.3 | 88.1 | 82.3 | 82.5 | 88.4 | 85.3 |
| RoBERTa | T5 | 87.1 | 89.0 | 87.9 | 82.8 | 90.2 | 86.3 |
| RoBERTa (Fine-tuned) | T5 | 85.8 | 90.3 | 88.0 | 87.5 | 90.1 | 88.8 |
| S-BERT | RoBERTa | 79.3 | 87.8 | 83.3 | 83.1 | 88.7 | 85.8 |
| S-BERT (Fine-tuned) | RoBERTa | 82.1 | 88.4 | 85.1 | 84.3 | 89.1 | 86.6 |
| S-BERT | T5 | 88.9 | 90.6 | 89.7 | 90.6 | 91.2 | 90.9 |
| S-BERT (Fine-tuned) | T5 | 89.2 | 92.5 | **90.8*** | 93.5 | 91.9 | **92.7*** |

Table 1: Performance comparison on the whole evaluation set. For the "Random" retriever, the negatives are randomly sampled to match our number of hard negatives. * designates statistical significance ($p < 0.05$).

| Method | ESC |  |  | RED |  |  | MAVEN-ERE |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | P | R | F1 | P | R | F1 | P | R | F1 |
| BiLSTM | 29.8 | 12.9 | 18.1 | 51.2 | 48.5 | 49.8 | 24.9 | 11.7 | 15.9 |
| BERT | 30.3 | 11.5 | 16.7 | 59.0 | 45.3 | 51.3 | 28.4 | 13.3 | 18.2 |
| RoBERTa | 31.9 | 14.4 | 21.5 | 61.3 | 48.7 | 54.3 | 28.6 | 12.7 | 17.6 |
| LIP | 36.2 | 23.5 | 28.2 | 64.8 | 57.6 | 61.0 | – | – | – |
| T5 | 34.8 | 26.7 | 30.2 | 64.2 | 54.6 | 59.0 | 27.4 | 23.5 | 25.3 |
| RichGCN | 36.4 | 32.1 | 34.1 | 68.9 | 60.2 | 64.3 | 34.4 | 20.5 | 25.7 |
| GPT-3.5 | 13.9 | 54.7 | 22.2 | 41.4 | 45.8 | 43.5 | – | – | – |
| ***Ours (Retriever + Classifier)** |  |  |  |  |  |  |  |  |  |
| RoBERTa + RoBERTa | 40.3 | 31.2 | 35.2 | 66.7 | 58.5 | 62.3 | 36.0 | 26.4 | 30.5 |
| S-BERT + RoBERTa | 40.6 | 34.2 | 37.1 | 76.9 | 59.5 | 67.1 | 36.2 | 26.9 | 32.0 |
| RoBERTa + T5 | 41.4 | 33.6 | 37.0 | 70.9 | 75.5 | **73.1** | 40.4 | 33.5 | 31.9 |
| S-BERT + T5 | 45.7 | 38.5 | **41.8*** | 87.4 | 62.1 | 72.6* | 36.8 | 29.5 | **32.8*** |

Table 2: Performance comparison on all non-negative event pairs with different retrievers and classifiers.

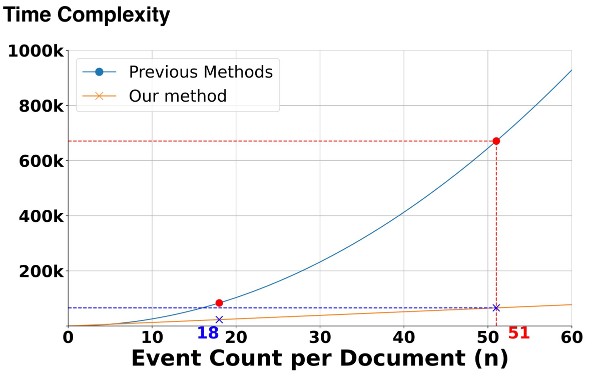

Figure 3: Inference time complexity comparison over *events per document* on ESC ($k = 5$).

Finally, after *fine-tuning* the retriever model, particularly S-BERT, the DERE performance can be further improved. Indeed, the fine-tuned retrievers significantly contribute to the overall performance and efficiency of DERE models. Our findings emphatically advocate for the integration of advanced retriever models as indispensable components of the DERE frameworks.

### 4.3 Additional Analysis

**Time Complexity Analysis.** For $m$ documents with $n$ events per document, conventional pairwise approaches exhibit a time complexity of $O(m * n^2)$. Our retriever narrows candidate pairs down to $k * n$ ($k$ candidate per event), and it scales linearly with matrix multiplication in inference. To quantify the efficiency gains of our method, we compare the time complexity at inference time. Figure 3 illustrates the quadratic growth versus our method's linear growth. Average *(n=18)* and maximal *(n=51)* event count is highlighted, in which our approach reduces approximately 70% at inference time.

| Retriever | Encoding | ESC |  |  | RED |  |  |
|---|---|---|---|---|---|---|---|
|  |  | P | R | F1 | P | R | F1 |
| **RoBERTa** | Trigger-only | 73.2 | 77.0 | 75.1 | 81.3 | 79.8 | 80.5 |
|  | Wrapped* | 85.8 | 90.3 | **88.0** | 87.5 | 90.1 | **88.8** |
|  | Graph-based | 72.1 | 86.1 | 78.8 | 82.5 | 81.4 | 81.9 |
| **S-BERT** | Trigger-only | 79.1 | 80.2 | 79.6 | 82.8 | 85.4 | 84.1 |
|  | Wrapped* | 89.2 | 92.5 | **90.8** | 93.5 | 91.9 | **92.7** |
|  | Graph-based | 79.1 | 81.4 | 80.2 | 83.2 | 91.5 | 87.2 |

Table 3: Performance comparison using different encoding strategies for both stages on the ESC dataset.

**Effectiveness of Encoding Strategies.** Table 3 shows a comparative analysis of various encoding strategies. Our *wrapped encoding*, as formulated in Equation 2, effectively aids models in recognizing

| $k$ | P | R | F1 |
|---|---|---|---|
| 3 | **91.4** | 67.2 | 77.5 |
| 5 | 89.2 | **92.5** | **90.8** |
| 7 | 85.0 | **93.2** | 88.9 |
| 10 | 79.9 | 91.8 | 85.4 |
| # EVENT | 86.1 | 85.7 | 85.9 |

Table 4: Relationship between the number of top event pairs retrieved in Stage 1 ($k$) and Stage 2 performance.

and processing the relevant information within a rich textual landscape, whereas *trigger-only encoding*, formulated as: $\mathbf{r}_h = Enc(\texttt{<m>}e_h\texttt{</m>})$, misses some contextual nuances. Surprisingly, *graph-based encoding* (Nguyen and Grishman, 2018) with syntactic dependency trees does not improve the performance, which might be attributed to the noise introduced due to its high complexity.

**Effect of Retrieved Candidate Count.** We further investigate the impact of the number of candidates retrieved per event ($k$), where S-BERT retriever and T5 classifier are used on the ESC dataset. The results are shown in Table 4, where *# EVENT* denotes without retrievers. When $k = 3$, high precision is offset by low recall, suggesting that too few event pairs limit relation detection. $k = 5$ offers the best performance, striking a balance between capturing relevant relations and avoiding classification overload. As $k$ increases beyond this point, while slowing the process by nature, more non-relevant pairs are also considered, making the classifier's training data more skewed as well, which detracts from the overall performance.

**Effectiveness of Coreference Chains.** Table 5 shows the experimental results after adding coreference chains information, which is defined as the event mentions referring to same events (Wang et al., 2022). The coreference chains are obtained from golden annotations and are incorporated as supplementary inputs. Experimental results show that the addition of coreference chains further enhance performance, regardless of which retriever is employed. Furthermore, the performance with the RoBERTa retriever gains more improvements, even outperforming S-BERT, possibly because RoBERTa are more proficient to leverage deep contextual insights from coreference chains.

### 4.4 Case Study

We further conduct a case study by sampling 50 event pairs that are mispredicted without the retriever but predicted correctly with the retriever model. We observe that the retriever model is particularly beneficial for *document-level* and *implicit* event relations because of the notable decrease in negative samples. As the following example:

> *A SAF spokesman denied the **attack** occurred. [...] did not explode, **fell** directly within the camp, [...]*

the events "***attack***" and "***fell***" span in separate sentences, and there are no causal clues (e.g., "cause" and "lead to") between them. Without the retrieval stage, the cross-encoders are unable to identify the relationship between them (i.e., ***attack*** is a precondition of ***fell***) because of the large proportion of negatives in the training set; however, with the cross-encoder trained on the samples retrieved by the retriever, the relationships between these samples are more likely to be recognized, alleviating the skewed distribution issue in DERE datasets.

| Model | P | R | F1 |
|---|---|---|---|
| Random (Retriever) | 86.1 | 85.7 | 85.9 |
| **RoBERTa (Retriever)** | 85.8 | 90.3 | 88.0 |
| + *Coref Chain* | 91.6 | **90.8** | **91.2** |
| **S-BERT (Retriever)** | 89.2 | 92.5 | 90.8 |
| + *Coref Chain* | **96.1** | 86.5 | **91.0** |

Table 5: Impact of coreference chains on ESC.

## 5 Conclusion and Future Work

We for the first time introduce a novel two-stage framework for DERE, which improves both efficiency and model training. It first uses a retriever to identify event pairs, then a cross-encoder for event relation prediction. Experimental results on three representative datasets underscore the effectiveness of our method, which significantly improves both accuracy and efficiency compared to the baseline models. We further investigate the efficacy of different encoding strategies, and demonstrate the effectiveness of leveraging coreference chains in candidate event pair identification. In the future, we will adapt our method to more IE tasks (e.g., entity relation extraction), and verify its generalizability.

## Limitations

Although our proposed two-stage framework performs well on DERE in terms of both overall performance and efficiency, it still has the following two limitations: (1) Due to the limitation of DERE datasets, we test the performance on the three most representative and wildly adopted datasets in other papers in this field. Future research demands the annotation of DERE datasets in other high-resource and low-resource languages to test the generalizability of our method. (2) For the second stage we employ the representative RoBERTa and T5 cross-encoders. More deliberated models or better prompting may yield better results, though they do not impact the conclusion of our experiments.

## References

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4862–4872, Online. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10804–10816, Toronto, Canada. Association for Computational Linguistics.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9451–9478, Bangkok, Thailand. Association for Computational Linguistics.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

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.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. *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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10901–10913, Toronto, Canada. Association for Computational Linguistics.

Xin Huang, Chor Seng Tan, Yan Bin Ng, Wei Shi, Kheng Hui Yeo, Ridong Jiang, and Jung Jae Kim. 2021. Joint generation and bi-encoder for situated interactive multimodal conversations. In *AAAI 2021 DSTC9 Workshop*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Ruosen Li, Zimu Wang, Son Quoc Tran, Lei Xia, and Xinya Du. 2024. Meqa: A benchmark for multi-hop event-centric question answering with explanations. In *Advances in Neural Information Processing Systems*, volume 37, pages 126835–126862. Curran Associates, Inc.

Li Lin, Yixin Cao, Lifu Huang, Shu'Ang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. 2022. What makes the story forward? inferring commonsense explanations as prompts for future event generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in*

*Information Retrieval*, SIGIR '22, page 1098–1109, New York, NY, USA. Association for Computing Machinery.

Jian Liu, Yubo Chen, and Jun Zhao. 2020a. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization. Main track.

Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. 2020b. Extracting events and their relations from texts: A survey on recent research progress and challenges. *AI Open*, 1:22–39.

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. *Preprint*, arXiv:1907.11692.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. 2022. MMEKG: Multi-modal event knowledge graph towards universal representation across modalities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 231–239, Dublin, Ireland. Association for Computational Linguistics.

Thien Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023a. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023b. OmniEvent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 508–517, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Haoyang Wen and Heng Ji. 2021. Utilizing relative event time to enhance event-event temporal relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10431–10437, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Baiyan Zhang, Qin Chen, Jie Zhou, Jian Jin, and Liang He. 2024. Enhancing event causality identification with rationale and structure-aware causal question answering. *Preprint*, arXiv:2403.11129.

Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023. Causal intervention for mitigating name bias in machine reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12837–12852, Toronto, Canada. Association for Computational Linguistics.

## A Baseline Models

We compare our method against various baselines: **BiLSTM** (Cheng and Miyao, 2017) captures the dependency paths between events. **BERT** (Devlin et al., 2019) and **RoBERTa** (Liu et al., 2019) are transformer-based discriminative models, and **T5** (Raffel et al., 2020) is a transformer-based generative model. **LIP** (Gao et al., 2019) combines document structure with textual content, identifying nuanced event relations using structural patterns. **RichGCN** (Tran Phu and Nguyen, 2021) employs Graph Convolutional Networks to create interaction graphs. Zhang et al. (2024) employs **GPT-3.5** (`turbo-1106`) to enhance zero-shot prediction. All baselines (GPT excluded) conduct pairwise comparisons.