# Dynamic Energy-Based Contrastive Learning with Multi-Stage Knowledge Verification for Event Causality Identification

**Ya Su[1], Hu Zhang[1,2]*, Yue Fan[1], Guangjun Zhang[1], Yujie Wang[1], Ru Li[1,2], Hongye Tan[1,2]**

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
su_ya6990@163.com, zhanghu@sxu.edu.cn, yuefan24@163.com, zgj2866@gmail.com, init_wang@foxmail.com, {liru,tanhongye}@sxu.edu.cn

## Abstract

Event Causal Identification (ECI) aims to identify fine-grained causal relationships between events from unstructured text. Contrastive learning has shown promise in enhancing ECI by optimizing representation distances between positive and negative samples. However, existing methods often rely on rule-based or random sampling strategies, which may introduce spurious causal positives. Moreover, static negative samples often fail to approximate actual decision boundaries, thus limiting discriminative performance. Therefore, we propose an ECI method enhanced by Dynamic Energy-based Contrastive Learning with multi-stage knowledge Verification (DECLV). Specifically, we integrate multi-source knowledge validation and LLM-driven causal inference to construct a multi-stage knowledge validation mechanism, which generates high-quality contrastive samples and effectively suppresses spurious causal disturbances. Meanwhile, we introduce the Stochastic Gradient Langevin Dynamics (SGLD) method to dynamically generate adversarial negative samples, and employ an energy-based function to model the causal boundary between positive and negative samples. The experimental results show that our method outperforms previous state-of-the-art methods on both benchmarks, EventStoryLine and Causal-TimeBank.

## 1 Introduction

Reasoning about causal relations between events is a core component of human language understanding. Event Causality Identification (ECI), a fundamental task in natural language processing (NLP), aims to determine whether a causal relationship exists between any two events within unstructured text. It has been widely applied in various scenarios such as knowledge graph construction (Chen et al., 2019; Khatib et al., 2020), question answering (Oh et al., 2017; Liu et al., 2023b), and event prediction (Radinsky et al., 2012).

Contrastive learning (Chao et al., 2024; Ding et al., 2024; Yin et al., 2023) has emerged as a promising approach for enhancing ECI by learning causal representations through the construction of positive and negative samples. However, existing methods still face challenges constructing high-quality training samples and accurately modeling causal boundaries. **Existing constructions of positive and negative samples suffer from insufficient precision:** most existing works predominantly rely on manual rules or random sampling to generate contrastive samples, which are often difficult to distinguish between semantically similar but causally distinct event pairs. Moreover, as shown in Table 1, causal instances are inherently scarce in widely used datasets such as EventStoryLine v9.0 (ESC) (Caselli and Vossen, 2017) and Cause-TimeBank (CTB) (Mirza and Tonelli, 2014). This scarcity further exacerbates the instability and noise susceptibility of contrastive learning during contrastive sample construction. Detailed analysis and discussion are provided in Appendix A.

**Spurious causal phenomena as a source of positive example noise:** current methods for constructing positive instances often rely on shallow features such as temporal order and lexical co-occurrence while neglecting deeper causal chains grounded in human reasoning, such as "psychological motives and results". This makes it difficult to accurately capture causal relationships under complex semantic contexts. As shown in Figure 1, event pairs such as $(fueled_{e5}, bashed_{e9})$ and $(bashed_{e9}, smashed_{e8})$ exhibit surface-level temporal or syntactic dependencies but lack direct causal relationships, serving as typical samples of spurious causality. Further analysis shows that introducing a mediating event can reveal an indirect causal chain $(fueled_{e5} \rightarrow riot_{e6} \rightarrow bashed_{e9})$, which tends to be overlooked and may introduce

*Corresponding author



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12249–12267
November 4-9, 2025 

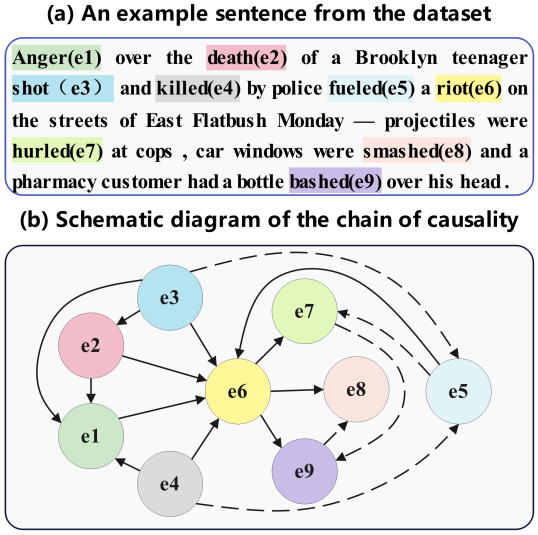


Figure 1: An example from the ESC dataset, where the dashed line represents spurious causal relationships, and the solid line represents genuine causal relationships.

noise into positive samples.

**Static negative samples cannot approximate the true decision boundaries:** current contrastive learning methods predominantly rely on static random negatives, resulting in a scattered negative distribution and uniform training signals that hinder the modeling of nearcausal boundary instance. In particular, when multiple event pairs with similar semantics but different causal attributes exist within the same sentence, simple negative sampling based on semantic similarity or lexical co-occurrence struggle to capture these subtle differences.

To improve the quality of contrastive samples in the ECI task, we propose a contrastive sample construction with a multi-stage knowledge verification, which integrates structured knowledge validation, generative causal reasoning, and multi-level semantic discrimination. We first build an initial candidate pool using multi-hop semantic associations derived from a heterogeneous event-concept graph. We then verify causal consistency through logical constraints from structured knowledge graphs and inference from generative knowledge models to eliminate spurious causal relations. Additionally, we employ a lightweight discriminator and a large language model (LLM) to further select and rank samples across multiple semantic levels, yielding more discriminative positive and negative samples.

To address the limitation of static negative samples in capturing true causal decision boundaries, we draw inspiration from the energy function's capacity to model node similarity in graph structures (Zeng et al., 2025), and propose a dynamic energy-based contrastive learning method based on SGLD (Welling and Teh, 2011). This method employs an adversarial dynamic sampling strategy to iteratively maximize the energy gap between positive and negative samples, thereby enhancing the modeling of causal boundaries in the embedding space. To summarize, we propose an ECI method enhanced by dynamic energy-based contrastive learning with multi-stage knowledge verification (DECLV). The main contributions of this paper are as follows:

- We propose a multi-stage verification mechanism that integrates multi-source knowledge verification and LLM-guided causal inference to generate high-quality contrastive samples and suppress spurious causal disturbances;
- We introduce a dynamic energy-based contrastive learning approach with SGLD to better distinguish causal boundaries between positive and negative samples;
- DECLV consistently outperforms SOTA baselines on two benchmarks, demonstrating superior causal relation identification.

## 2 Related Work

With the development of large language models (LLMs), ECI has shifted from shallow pattern-based methods to deep semantic reasoning frameworks. Early research primarily relied on pattern matching and feature engineering, constructing classifiers based on lexical, syntactic, and statistical features (Beamer and Girju, 2009). As deep learning techniques advanced, researchers have explored four major directions to enhance ECI: **Knowledge Augmentation,** which integrates external knowledge bases such as ConceptNet (Speer et al., 2017) and WordNet (Miller, 1995) to compensate for the absence of explicit causal clues (Zuo et al., 2021b; Chen and Mao, 2024; Ding et al., 2024; Wu et al., 2023); **Graph-based Reasoning,** which leverages graph neural networks to capture event interactions and dependencies (Pu et al., 2023), such as event interaction graphs (Tran Phu and Nguyen, 2021) and semantic structures (Hu et al., 2023); **Prompt-based Learning,** which adapts causal inference via task-specific prompts, such as combining knowledge injection (Liu et al., 2023a) and event con-

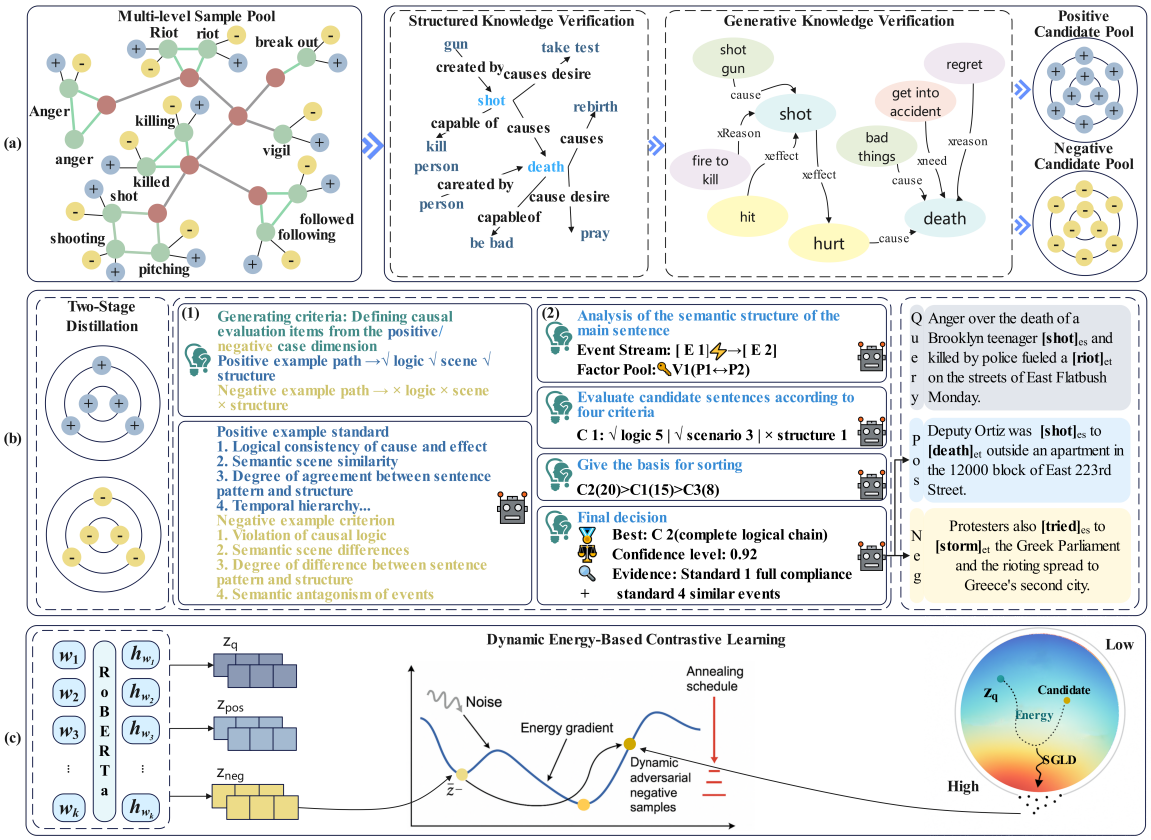


Figure 2: Overall framework of the proposed DECLV method: (a) multi-source knowledge verification for contrastive sample generation; (b) LLM-guided causal verification for sample selection and ranking; (c) dynamic energy-based contrastive learning with SGLD.

| Metric | Event StoryLine | Cause-TimeBank |
|---|---|---|
| Number of event-mention pairs | 54,326 | 9,631 |
| Causal proportion among event-mention pairs | 10.36% | 6.7% |
| Number of event-concept pairs | 34,491 | 7,608 |
| Causal proportion among event-concept pairs | 5.26% | 4.18% |

Table 1: Statistics of causal ratios for event mention and concept pairs in ESC and CTB.

cept heterogeneous graphs (Su et al., 2025) to enhance semantic alignment; **Contrastive Learning,** which strengthens event pair discrimination by introducing positive and negative samples, such as context-aware contrastive learning (Yin et al., 2023; Chao et al., 2024) and knowledge graph-enhanced contrastive learning (Ding et al., 2024). These advancements reflect a clear trend: from shallow, manually engineered features toward multidimensional semantic modeling that integrates external knowledge, structural reasoning, prompting, and contrastive learning-laying the foundation for more robust and interpretable ECI systems.

## 3 Methodology

### 3.1 Task Definition

Given a document containing multiple sentences $D = \{s_1, s_2, \ldots, s_n\}$, each with several events $E = \{e_1, e_2, \ldots, e_m\}$, as shown in Figure 1(b), the goal of ECI is to determine whether a causal relation exists between any two events $(e_s,\ e_t)$ $(s \neq t)$ in the event set $E$ based on contextual and semantic cues. Figure 2 illustrates the framework of our proposed DECLV approach, which comprises three main modules: (a) multi-source knowledge verification (Section 3.2) and (b) LLM-guided causal verification (Section 3.3), which together constitute a multi-stage knowledge verification strategy; and (c) dynamic energy-based contrastive learning with

SGLD (Section 3.4). We describe each component in detail below.

### 3.2 Contrastive Sample Construction with Multi-Source Knowledge Verification

**Multi-level Sample Pool.** To enhance the diversity of positive and negative samples in the ECI task, we propose a multi-level sample pool construction mechanism based on an event concept heterogeneous graph (Su et al., 2025). As shown in Figure 2(a), a three-layer structure *event concept – event mention – candidate sample* is designed to provide semantic grounding and contextual linkage. If an event mention node lacks available candidate pairs, it inherits the samples from its corresponding event concept node. Moreover, if multi-hop connections exist between event mentions under the same event concept, their associated samples are collected via dependency paths to further enrich the semantic coverage of the sample pool.

**Causality Verification.** While existing studies often use structured knowledge bases like ConceptNet for knowledge augmentation, ATOMIC (Sap et al., 2019) focuses more on causal reasoning behind everyday human behavior, yet remains underutilized in ECI tasks. The (COMET-) ATOMIC 2020 (Hwang et al., 2021) model is primarily trained on ATOMIC, with supplementary knowledge from ConceptNet, and focuses on causal knowledge generation, exhibiting enhanced capabilities in causal reasoning. Therefore, as shown in Figure 2(a), we design a multi-source causality verification module that integrates structured knowledge, generative commonsense reasoning, and semantic similarity computation, comprising the following stages:

(1) Structured Knowledge Verification. For the input event pair $(e_s,\ e_t)$, explicit causal edges between them are extracted from ConceptNet, including the *Causes* and *LeadsTo* relations. If a direct connection exists, the event pair is classified as a high-confidence positive sample; If no direct connection is found, the word-level similarity from WordNet (Miller, 1995) and phrase-level similarity from SBERT (Reimers and Gurevych, 2019) are further used to attempt matching with other expressions in ConceptNet for validation.

(2) Generative Knowledge Reasoning. If structured verification fails to confirm the causal relation between an event pair $(e_s,\ e_t)$, we employ the (COMET-) ATOMIC 2020 model for generative reasoning verification. Specifically, the model generates diverse causal inferences centered on $e_s$, and each inferred event is scored based on its semantic similarity to $e_t$ using SBERT and Word2Vec (Mikolov et al., 2013). Pairs with low overall similarity are treated as spurious causal samples and added to the negative pool. Details on the reasoning types, matching rules, and distillation criteria are provided in Appendix B. This process leverages ATOMIC's "psychological motives and results" causal chains, effectively compensating for the lack of subjective reasoning in structured knowledge and improving the identification and verification of complex causal paths between events.

### 3.3 LLM-Guided Causal Verification

**Lightweight Model-based Preliminary Distillation.** In this stage, we employ lightweight PLMs to efficiently perform and coarse-to-fine distillation of candidate samples, aiming to construct a low-noise, high-discriminative candidate set, as shown in Figure 2(b).

(1) MPNet-based Initial Distillation. Given a query sample $q$, MPNet (Song et al., 2020) encodes its sentence representation $h_q$. Then, we separately retrieve the Top-$K$ most similar samples from the positive candidate pool and the negative candidate pool based on cosine similarity, forming the initial candidate positive set $\mathcal{C}_q^{pos}$ and candidate negative set $\mathcal{C}_q^{neg}$. The similarity is computed as:

$$h_i = \mathrm{MPNet}(x_i), \quad x_i \in \mathrm{CandidatePool} \tag{1}$$

$$\mathrm{sim}(\mathrm{h}_q, \mathrm{h}_i) = \frac{\mathrm{h}_q \cdot \mathrm{h}_i}{\|\mathrm{h}_q\|\|\mathrm{h}_i\|} \tag{2}$$

(2) Semantic Matching with SimCSE. To enhance robustness, SimCSE (Gao et al., 2021), a contrastive learning-based sentence encoder, is used to re-encode the query and each candidate sample into deep semantic vectors $s_q$ and $s_i$. Based on the refined similarity scores, we separately re-rank $\mathcal{C}q^{pos}$ and $\mathcal{C}q^{neg}$. To unify the selection process, we define a selection operator $\mathrm{Select}_{K/2}(\cdot)$: given a candidate set, it selects $K/2$ samples according to the scoring rule controlled by a sign variable $\sigma_l$. Specifically, if $l = \mathrm{pos}$, then $\sigma_l = +1$, indicating that the top-$K/2$ most similar positive samples to the query are retained; if $l = \mathrm{neg}$, then $\sigma_l = -1$, meaning that the bottom-$K/2$ least similar negative samples are preserved, as defined in Equations 3 and 4.

$$\mathbf{s}_i = \mathrm{SimCSE}(x_i) \tag{3}$$

$$\hat{\mathcal{C}}_q^l = \mathrm{Select}_{K/2}\big(\{\, \sigma_l \cdot sim(s_q, s_i) \mid x_i \in \mathcal{C}_q^l \}\big) \tag{4}$$

Finally, two subsets are formed: the positive sample set $\mathcal{P}_q$, which includes candidate positive samples that are semantically consistent with the causal logic of $q$, and the negative sample set $\mathcal{N}_q$, which contains candidate negative samples that are semantically irrelevant and non-causal.

**LLM-Guided Fine-grained Selection and Ranking.** Building on initial distillation by lightweight models, we introduce DeepSeek LLM (Bi et al., 2024) to perform causal reasoning and fine-grained ranking over candidate samples. Its optimized sparse attention enables effective modeling of long-range event dependencies, while its causality-oriented pretraining (DeepSeek-AI et al., 2024) better aligns with the needs of ECI tasks. As shown in Figure 2(b1), we employ prompt engineering to guide the LLM in multi-dimensional causal evaluation of candidate pairs (see Appendix F), and perform ranking and selection based on aggregated scores (see Appendix G). In Figure 2(b2), the top-scoring positive and lowest-scoring negative sample is retained based on the multi-dimensional scores, as defined in Equations 5 and 6.

$$p_q = \arg\max \mathrm{Score}(x) \quad x \in \mathcal{P}_q \tag{5}$$

$$n_q = \arg\min \mathrm{Score}(x) \quad x \in \mathcal{N}_q \tag{6}$$

Finally, based on the selected samples, we employ RoBERTa (Liu et al., 2019) as the encoder to integrate prompt information during sentence embedding and construct training triplets $(z_q,\ z_{pos},\ z_{neg})$. We design an analogy-style prompt template and a causal label set for the ECI task to compute the causal scores of event pairs, as detailed in Appendix C. In this way, the prompt information is effectively incorporated into the event embeddings, enhancing both the model's causal perception ability and the semantic discriminability of the samples. The proposed two-stage distillation framework combines the efficiency of lightweight models with the causal reasoning capabilities of the LLM, providing high-quality supervision for subsequent energy-based contrastive learning.

### 3.4 Dynamic Energy-Based Contrastive Learning with SGLD

**Energy Function Modeling.** To capture deeper and dynamically evolving semantic interactions between query samples and candidate samples, we design an energy function based on multi-dimensional feature fusion, as defined in Equation 7. As shown in Figure 2(c), this function assigns low energy values to causal event pairs and high energy values to non-causal event pairs, progressively widening the semantic gap between positive and negative samples in the embedding space, thereby enhancing the discriminative power of the representations. In each input group, the query sample $z_q$ represents the encoded sentence containing a specific event pair with prompt augmentation, while the candidate samples $z \in Z$, where $Z = z_{pos} \cup z_{neg}$, include both positive and negative samples. Specifically, the input to the energy function is composed of five components: the concatenation $z_q$ and $z$ (i.e., $[z_q; z]$), their element-wise interaction $z_q \odot z$, the difference vector $z_q - z$, and a semantic attention vector generated by a gating mechanism, defined as $\sigma(W_g[z_q; z] + b_g)$, where $W_g$ and $b_g$ are learnable parameters, and $\sigma$ denotes the sigmoid activation function.

$$E(z_q, z) = -W_2 \cdot \mathrm{ReLU}\Big(W_1 \cdot \big[z_q; z; z_q \odot z; z_q - z; \sigma\big(W_g \cdot [z_q; z] + b_g\big)\big] + b_1\Big) + b_2 \tag{7}$$

After feature fusion, the resulting representation is fed into the first linear layer with weight matrix $W_1 \in \mathbb{R}^{h \times 5d}$ and bias $b_1 \in \mathbb{R}^h$. The ReLU-activated output is then passed through a second linear layer with weight matrix $W_2 \in \mathbb{R}^{1 \times h}$ and bias $b_2 \in \mathbb{R}$ to produce the final energy value. The energy function outputs a scalar value that reflects the semantic compatibility between $z_q$ and $z$, where a lower energy value indicates a stronger causal correlation. By introducing a negative sign, the model is encouraged during training to lower the energy values of positive samples and raise those of negative samples, thereby learning a sharper causal decision boundary in the embedding space.

**Dynamic Negative Sample Generation via SGLD.** In contrastive learning, the quality of sample construction directly affects the discriminative capability of the model. Since positive samples typically exhibit clear semantic features but are limited in quantity, excessive augmentation may

lead the model to overfit superficial patterns, making it challenging to learn intrinsic representations of causal structures. Therefore, improving sample quality hinges on optimizing the construction of negative samples. To this end, we introduce the Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011) method, which iteratively generates more semantically challenging adversarial negative samples by injecting noise and incorporating energy gradient information based on the original negative samples. This approach dynamically enhances the discriminative capacity of negative samples and employs an annealing mechanism to facilitate progressive exploration of the energy space. Specifically, the update rule for SGLD is as follows.

$$\tilde{z}^{(t+1)} = \tilde{z}^{(t)} - \frac{\eta_t}{2}\nabla_{\tilde{z}} E(z_q, \tilde{z}^{(t)}) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \eta_t) \tag{8}$$

Here, $z_q$ denotes the query sample, $\tilde{z}^{(t)}$ represents the negative sample generated at the $t$-th iteration, $\mathcal{N}_t$ is the annealing step size at step $t$, and $\epsilon_t$ is the Gaussian noise term. A polynomial annealing strategy is employed to control the variation of the step size.

$$\eta_t = \eta_{end} + (\eta_{start} - \eta_{end})\left(1 - \frac{t+1}{T}\right)^{\gamma} \tag{9}$$

This strategy allows for greater exploratory capability during the early sampling stages while gradually stabilizing in the later stages. It facilitates the generation of adversarial negative samples that lie closer to the semantic boundary, thereby enhancing the model's sensitivity and discriminative power near the decision boundary.

$$\mathcal{L}_{\text{con}}(z_q) = -\log \frac{e^{-E(z_q, z_{\text{pos}})/\tau}}{e^{-E(z_q, z_{\text{pos}})/\tau} + e^{-E(z_q, \tilde{z}_{\text{neg}})/\tau}} \tag{10}$$

Finally, we use the InfoNCE contrastive loss function for training, as shown in Equation 10. Here, $\tilde{z}_{\text{neg}}$ denotes the dynamic negative samples generated by SGLD, and $\tau$ is the temperature coefficient. This loss function strengthens the energy difference between positive and negative samples, guiding the model to form a clearer causal semantic boundary in the embedding space.

### 3.5 Training strategy

For each sentence $S$ in the input document $D^*$ and its annotated event pairs $(e_s,\ e_t)$, we construct the input based on a prompt template. The causal relationship between the event pair is determined by predicting the probability of the word at the `[MASK]` position. The model's output probability at the `[MASK]` position is denoted as $p_{st} \in (0, 1)$, while the ground-truth label $y_{st} \in (0, 1)$ indicates whether the event pair exhibits a causal relationship. We supervise the prediction results using the following cross-entropy loss function.

$$\mathcal{L}_{\text{ce}}(e_s, e_t) = - [y_{st} \log p_{st} + (1 - y_{st}) \log(1 - p_{st})] \tag{11}$$

Finally, we combine the cross-entropy loss and the contrastive loss with appropriate weighting to form the complete training objective.

$$\mathcal{L}_{\text{total}}(e_s, e_t) = \mathcal{L}_{\text{ce}}(e_s, e_t) + \lambda \cdot \mathcal{L}_{\text{con}}(z_q) \tag{12}$$

Here, $\lambda$ is a hyperparameter used to balance the causal prediction and semantic contrastive tasks.

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

We conduct evaluations on two publicly available ECI benchmark datasets: EventStoryLine v9.0 (ESC) (Caselli and Vossen, 2017) and Causal-TimeBank (CTB) (Mirza and Tonelli, 2014), with their statistics summarized in Appendix A: ESC consists of 22 topics, 258 documents, and 5,334 event mentions, while CTB contains 183 documents and 6,811 event mentions. Following widely adopted data-splitting strategies from previous studies (Gao et al., 2023; Chao et al., 2024; Su et al., 2025), we apply 5-fold document-level cross-validation on ESC and 10-fold on CTB. In addition, we adopt Precision (P), Recall (R), and F1-score (F1) as evaluation metrics.

### 4.2 Parameter Settings

We adopt the pre-trained language model RoBERTa-base[1] as the backbone encoder to perform contextual modeling over the input event sequences. The model consists of 12 Transformer encoder layers, each with a hidden size of 768 and

[1]https://huggingface.co/roberta-base/

| Methods | Model | EventStoryLine | | | Cause-TimeBank | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| Feature-based methods | DD(Mirza and Tonelli, 2014) | - | - | - | 67.3 | 22.6 | 33.9 |
| | Seq (Choubey and Huang, 2017) | 32.7 | 44.9 | 37.8 | - | - | - |
| Knowledge-augmented methods | LearnDA(Zuo et al., 2021b) | 42.2 | 69.8 | 52.6 | 41.9 | 68 | 51.9 |
| | KADE(Cao et al., 2021) | 61.5 | 73.2 | 66.8 | 56.8 | 70.6 | 66.7 |
| | DPF(Huang et al., 2024) | 55.9 | 69.8 | 62.1 | 53.7 | 64.2 | 58.5 |
| Graph-based methods | RichGCN(Tran Phu and Nguyen, 2021) | 49.2 | 63.0 | 55.2 | 39.7 | 56.5 | 46.7 |
| | SemSIn(Hu et al., 2023) | 50.5 | 63.0 | 56.1 | 52.3 | 65.8 | 58.3 |
| | ECLEP(Pu et al., 2023) | 49.3 | 68.1 | 57.1 | 50.6 | 63.4 | 56.3 |
| Prompt-adjusted methods | KEPT(Liu et al., 2023a) | 50.0 | 68.8 | 57.9 | 48.2 | 60 | 53.5 |
| | DPJL(Shen et al., 2022) | 65.3 | 70.8 | 67.9 | 63.6 | 66.7 | 64.6 |
| | LCKER(Su et al., 2025) | 65.3 | 70.8 | 67.9 | 63.6 | 66.7 | 64.6 |
| Contrastive learning methods | CauSeRL(Zuo et al., 2021a) | 41.9 | 69.0 | 52.1 | 43.6 | 68.1 | 53.2 |
| | GCKAN(Ding et al., 2024) | 50.9 | 60.6 | 55.3 | 52.2 | 60.7 | 56.1 |
| | ICCL(Chao et al., 2024) | 67.5 | 73.7 | 70.4 | 63.7 | 68.8 | 65.4 |
| | **DECLV(ours)** | **71.8** | **74.3** | **73.8** | **67.8** | **71.3** | **72.4** |

Table 2: Experimental Results on the ESC and CTB Datasets(%).

12 attention heads. During training, we employ AdamW (Loshchilov and Hutter, 2017) as the optimization algorithm. The learning rate is set to 1e-5 for the pre-trained parameters and 1e-4 for the newly added parameters. The batch size is set to 6, and the model is trained for 20 epochs. For structured knowledge validation, the word-level similarity threshold is set to 0.6, and the sentence-level threshold is set to 0.75. For the generative reasoning module, the maximum length of generated text is 50, with up to 5 candidate events. In the lightweight distillation stage, we adopt a top-k (k=100) strategy to filter high-confidence candidates. In the contrastive learning module, the weight of the contrastive loss $\lambda$ is set to 0.5, and the temperature parameter $\tau$ is set to 0.07.

### 4.3 Baselines

To demonstrate the effectiveness of this work, we compare our method with previous state-of-the-art models. For the EventStoryLine and Cause-TimeBank datasets, the baselines include the following five categories. Feature-based methods: **Seq** (Choubey and Huang, 2017), **DD** (Mirza and Tonelli, 2014). Knowledge-augmented methods: **LearnDA** (Zuo et al., 2020), **KADE** (Wu et al., 2023), **DPF** (Huang et al., 2024). Graph neural network-based methods: **RichGCN** (Tran Phu and Nguyen, 2021), **SemSIn** (Hu et al., 2023), **ECLEP** (Pu et al., 2023). Prompt-adjusted methods: **KEPT** (Liu et al., 2023a), **DPJL** (Shen et al., 2022), **LKCER** (Su et al., 2025). Contrastive learning methods: **CLINK** (Zuo et al., 2021a), **GCKAN** (Ding et al., 2024), **ICCL** (Chao et al., 2024). an in-context contrastive learning approach for ECI. A detailed description of the baselines is provided in Appendix D.

### 4.4 Main Result

Table 2 presents our experimental results on the ESC and CTB datasets, clearly demonstrating the performance differences among various methods on the ECI task. Overall, knowledge-enhanced and graph-based reasoning approaches alleviate the limitations of feature-driven models in event representation by incorporating external knowledge and modeling event structures. Prompt-based learning methods further exploit latent causal knowledge within PLMs by reformulating causal prediction as a PLM-driven lexical generation task. Building on these foundations, the integration of contrastive learning, as exemplified by methods such as ICCL, further enhances the model's ability to distinguish between causal and non-causal event pairs. Notably, our proposed DECLV method surpasses existing state-of-the-art models by 3.4% and 7% in F1 score on the ESC and CTB datasets, respectively. Its superior performance on the low-resource CTB dataset further demonstrates the effectiveness of our sample optimization strategy in scenarios with limited annotated data. Further analysis reveals that DECLV mitigates the noise typically introduced

| Methods | EventStoryLine | | | Cause-TimeBank | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| DECLV | **71.8** | **74.3** | **73.8** | **67.8** | **71.3** | **72.4** |
| -ECL | 71.3 | 67.1 | 69.5 | 66.9 | 70.5 | 70.9 |
| -QSS | 70.9 | 66.7 | 66.5 | 65.5 | 66.1 | 65.8 |
| -$CL_{da}$ | 69.0 | 65.9 | 67.2 | 66.6 | 67.1 | 66.4 |
| -$CL_{qo}$ | 65.4 | 68.8 | 66.7 | 65.2 | 67.4 | 65.8 |

Table 3: Ablation Results of Different Modules on the ESC and CTB Datasets (%).

during sample construction by incorporating high-quality positive and negative samples. Moreover, the dynamically generated negative samples effectively reshape the semantic space, encouraging the model to establish clearer causal boundaries.

### 4.5 Ablation Study

This section analyzes the contribution of each module in DECLV to performance, as well as the design of the energy function, through ablation studies.

**Module Ablation.** As shown in Table 3, we evaluate the following settings: **-ECL**, which removes the SGLD-based energy contrastive learning while retaining static samples to assess the impact of dynamic semantic perturbation; **-QSS**, which replaces the generation and selection mechanism with random similarity sampling to evaluate the effect of sample quality; **-$CL_{da}$**, which retains high-quality samples but removes the contrastive loss to isolate the effect of data augmentation; and **-$CL_{qo}$**, which uses only original query samples to assess prompt learning performance without any contrastive information. The results show that each module contributes significantly to modeling causal-semantic boundaries. Notably, the SGLD mechanism and high-quality sample strategy provide the greatest improvements in both discriminative ability and generalization performance.

**Feature Ablation of the Energy Function.** To further validate the effectiveness of each input feature in the energy function, we conduct feature ablation experiments by removing them one at a time: **-Interaction**, which removes element-wise interactions $z_q \odot z$ to evaluate the contribution of fine-grained semantic alignment to causal compatibility; **-Difference**, which removes difference vectors $z_q - z$ to assess the role of directionality and relative displacement in causal discrimination; and **-Attention**, which removes gated attention mechanism $\sigma\big(W_g \cdot [z_q; z] + b_g\big)$ to examine whether adaptive dimension selection helps focus on discriminative subspaces. As shown in Table 4, each feature contributes positively to model performance, with element-wise interaction and gated attention showing the most significant impact on causal discrimination ability.

| Setting | Removed Feature | P | R | F1 |
|---|---|---|---|---|
| Full model | – | 67.8 | 71.3 | 72.4 |
| -Interaction | $z_q \odot z$ | 61.9 | 70.6 | 69.7 |
| -Difference | $z_q - z$ | 65.7 | 69.5 | 72.2 |
| -Attention | Gated attention | 65.7 | 64.5 | 67.3 |

Table 4: Feature ablation of the energy function on the CTB dataset (%).

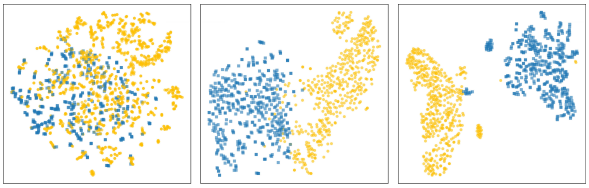


Figure 3: Embedding visualization of event pairs on ESC corpus. The yellow dots represent correctly predicted positive samples, and the blue dots represent correctly predicted negative samples.

### 4.6 Embedding Visualization

To verify the effectiveness of the DECLV method, we visualize the embedding distributions on the ESC test set using t-SNE (Hinton and Roweis, 2002), as shown in Figure 3. *w/o CL* denotes prompt learning without contrastive learning, and *Standard CL* refers to standard contrastive learning with similarity-based sampling. The rightmost figure shows our method. Compared to the baselines, DECLV yields clearer boundaries between positive and negative samples, effectively enhancing the model's ability to distinguish boundary events.

### 4.7 Case Study

This section presents a case study to demonstrate the effectiveness of the DECLV method. Due to the class imbalance between positive and negative samples in the ESC and CTB datasets, random sampling often results in positive examples that are overly similar to the query sentence, thereby weakening the training signal. As shown in Figure 4, ICCL tends to select positive examples with high similarity to the query sentence, leading to limited semantic diversity. In contrast, DECLV generates and selects semantically relevant but diverse positive and negative samples, thereby improving the

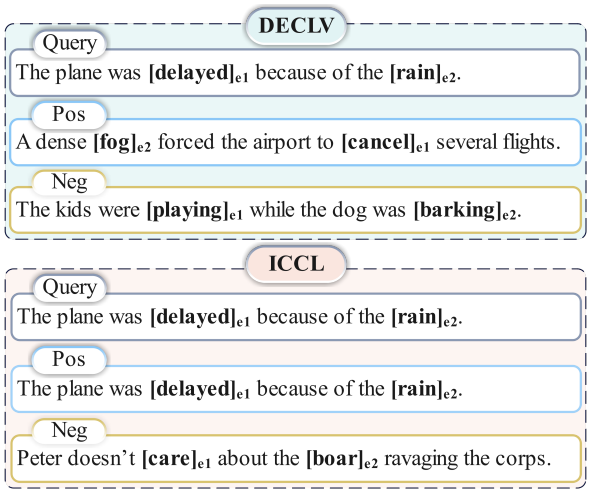


Figure 4: Case Study. Comparison of semantic diversity in positive and negative samples generated by DECLV (Top) and ICCL (Bottom).

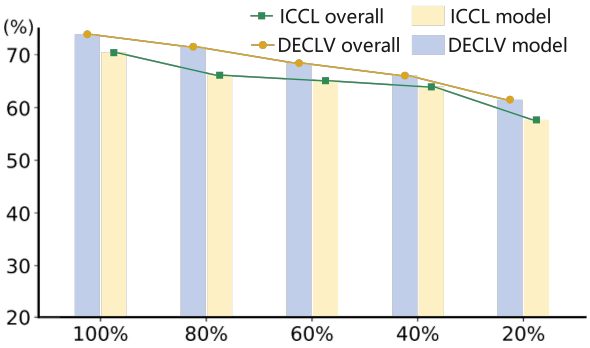


Figure 5: Results of few shot on ESC corpus.

modeling and identification of causal boundaries while preserving causal relevance.

### 4.8 Few-shot Setting

We compare the F1 performance of DECLV and ICCL under varying proportions of training data on the ESC dataset, as shown in Figure 5. As the training data decreases, DECLV shows a substantially smaller performance drop compared to the baseline. For example, when only 20% of the training data is retained, DECLV still significantly outperforms ICCL, demonstrating superior generalization in few-shot settings.

Furthermore, under the zero-shot setting, we compare the sentence-level ECI performance of PLMs such as BERT and RoBERTa, and mainstream LLMs including GPT-3.5-turbo, GPT-4 (Goswami et al., 2020), and DeepSeek-V3 (DeepSeek-AI et al., 2024). The results, summarized in Table 5, demonstrate that DECLV consistently outperforms both PLMs and LLMs in fine-grained local causal inference. A more detailed analysis and discussion of these findings are provided in Appendix E.

| Methods | EventStoryLine | | | Cause-TimeBank | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| BERT | 38.1 | 56.8 | 45.6 | 41.4 | 45.8 | 43.5 |
| RoBERTa | 42.1 | 64.0 | 50.8 | 39.9 | 60.9 | 48.2 |
| gpt-3.5-turbo | 27.6 | 80.2 | 41.0 | 6.9 | 82.6 | 12.8 |
| gpt-4 | 27.2 | 94.7 | 42.2 | 6.1 | 97.4 | 11.5 |
| DeepSeek-V3 | 20.9 | 74.5 | 30.3 | 3.5 | 23.3 | 5.91 |

Table 5: Intra-sentence causality identification results of different PLMs and LLMs on the ESC and CTB corpus.

## 5 Conclusion

We propose a dynamic energy-based contrastive learning method for ECI. The approach integrates structured knowledge validation, generative causal reasoning, and multi-level semantic discrimination capabilities of the LLM to construct high-quality training instances and mitigate spurious causal noise. Furthermore, we introduce an SGLD-based mechanism for dynamic adversarial negative sampling and employs an energy function to model the causal boundary between positive and negative samples. The experimental results on two widely used datasets demonstrate the effectiveness of our method in improving ECI performance.

## Limitations

In this paper, we investigate the role of a single energy function in modeling positive and negative samples, without exploring multi-level energy interactions(Deng et al., 2023; Zhang et al., 2024). The current energy function also lacks adaptivity and relies on manually defined structures and parameters, limiting its generalization. While it captures causal boundaries numerically, its semantic interpretability remains limited. Future work could incorporate explainability-oriented analysis methods (Fan et al., 2025; Zhao et al., 2024; Fan et al., 2024) to further explore the semantic nature of causal boundaries.

## Acknowledgements

We thank all the anonymous reviewers for their constructive comments and suggestions. This work is supported by the National Natural Science Foundation of China (62176145, 62476161), the Major Programs of the National Social Science Fund of China (24&ZD227), and the Interdisciplinary Research Fund of Shanxi University.

## References


Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09, page 430441, Berlin, Heidelberg. Springer-Verlag.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, and 67 others. 2024. Deepseek LLM: scaling open-source language models with longtermism. *CoRR*, abs/2401.02954.

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.

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

Siyuan Chen and Kezhi Mao. 2024. Explicit and implicit knowledge-enhanced model for event causality identification. *Expert Systems with Applications*, 238:122039.

Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. UHop: An unrestricted-hop relation extraction framework for knowledge-based question answering. 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 345–356, Minneapolis, Minnesota. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1796–1802. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Shumin Deng, Shengyu Mao, Ningyu Zhang, and Bryan Hooi. 2023. SPEECH: Structured prediction with energy-based event-centric hyperspheres. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–363, Toronto, Canada. Association for Computational Linguistics.

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

Yue Fan, Hu Zhang, Ru Li, YuJie Wang, Hongye Tan, and Jiye Liang. 2024. FRVA: Fact-retrieval and verification augmented entailment tree generation for explainable question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9111–9128, Bangkok, Thailand. Association for Computational Linguistics.

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

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? A comprehensive evaluation. *CoRR*, abs/2305.07375.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Ankur Goswami, Akshata Bhat, Hadar Ohana, and Theodoros Rekatsinas. 2020. Unsupervised relation extraction from language models using constrained cloze completion. *CoRR*, abs/2010.06804.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

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.

Peixin Huang, Xiang Zhao, Minghao Hu, Zhen Tan, and Weidong Xiao. 2024. Distill, fuse, pre-train: Towards effective event causality identification with commonsense-aware pre-trained model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 5029–5040. ELRA and ICCL.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Khalid Al Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7367–7374. AAAI Press.

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

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

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

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

T Mikolov, K Chen, G Corrado, and J Dean. 2013. International conference on learning representations. In *Efficient Estimation of Word Representations in Vector Space*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

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, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Ruili Pu, Yang Li, Suge Wang, Deyu Li, Jianxing Zheng, and Jian Liao. 2023. Enhancing event causality identification with event causal label and event pair interaction graph. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10314–10322, Toronto, Canada. Association for Computational Linguistics.

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

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2288–2299. International Committee on Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

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

Max Welling and Yee Whye Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681688, Madison, WI, USA. Omnipress.

Sifan Wu, Ruihui Zhao, Yefeng Zheng, Jian Pei, and Bang Liu. 2023. Identify event causality with knowledge and analogy. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

Maoxin Yin, Yuheng Chen, Huixun Qian, Haifeng Liu, and Junsheng Zhou. 2023. Enhancing chinese event causality identification with supervised contrastive learning. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 265–271.

Xianlin Zeng, Yufeng Wang, Yuqi Sun, Guodong Guo, Wenrui Ding, and Baochang Zhang. 2025. Graph structure refinement with energy-based contrastive learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 22326–22335. AAAI Press.

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

Yunxiao Zhao, Zhiqiang Wang, Xiaoli Li, Jiye Liang, and Ru Li. 2024. AGR: Reinforced causal agent-guided self-explaining rationalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 510–518, Bangkok, Thailand. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021a. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2162–2172. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3558–3571, Online. Association for Computational Linguistics.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. Knowdis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1544–1550. International Committee on Computational Linguistics.

## A Analysis of Causal Event Sparsity in Datasets

To further investigate the data-level challenges faced by current contrastive learning methods in ECI, we conducted a statistical analysis on two widely used datasets: EventStoryLine v9.0 (ESC) (Caselli and Vossen, 2017) and Cause-TimeBank (CTB) (Mirza and Tonelli, 2014). Specifically, we calculated the proportion of causal relations among event-mention pairs and event-concept pairs. As shown in Table 1, the proportion of causal event-mention pairs is only 10.36% in ESC and even lower in CTB, at 6.7%. When aggregating these mentions into higher-level event-concept pairs (Su et al., 2025), the proportions drop significantly to 5.26% and 4.18%, respectively. This analysis highlights the difficulty of obtaining high-quality positive examples from the original datasets and explains why contrastive learning approaches are particularly vulnerable to data sparsity and noise during sample construction.

## B Generative Causal Reasoning Distillation Rules

In order to more accurately verify the causal relationship between an event pair, we apply the (COMET-)ATOMIC2020 (Hwang et al., 2021) commonsense reasoning model when structured knowledge verification fails to identify a direct causal link. The model generates candidate events associated with the source event $e_s$, covering five types of causal relations: *Causes, xEffect, xReason, xIntent,* and *xNeed*. We assign different weights to these relations: *Causes, xReason,* and *xEffect*, which indicate more direct causal links, are assigned greater weight, while *xIntent* and *xNeed*, which represent motivational or prerequisite relations, are treated as auxiliary cues.

We then compute the semantic similarity between each generated candidate and the target event $e_t$ using SBERT (Reimers and Gurevych, 2019) and Word2Vec (Mikolov et al., 2013), and assign scores accordingly. If the overall score falls below a predefined threshold, the event pair is considered a spurious causal relationship and added to the negative sample pool. The main distillation rules include:

- $e_s$ and $e_t$ share the same xReason, suggesting a common cause rather than a direct causal link;
- $e_s$ is the xReason of $e_t$ but lacks an explicit Causes relation;
- The xEffect inferred from $e_s$ is semantically dissimilar to $e_t$.

By incorporating COMET-based generative reasoning and leveraging the psychologically grounded "psychological motives and results" chains provided by the ATOMIC knowledge base, this approach effectively compensates for the limitations of structured knowledge, which focuses on objective facts and lacks subjective causal reasoning. It enhances the verification of causal plausibility between event pairs and improves the modeling and validation of complex causal pathways.

## C Prompt Design for PLM Encoding

To enhance the causal awareness of PLMs during the embedding stage, we introduce a prompt-based learning mechanism that concatenates the event pairs in the query sample with an analogy-style prompt template, guiding the model to focus on causal relation discrimination. The prompt template is shown in Equation 13. We adopt RoBERTa (Liu et al., 2019) as the base encoder, where event semantics and task instructions are integrated into the model input via prompting.

$$\begin{aligned} ECI(q) = &\, The\ event < /t_1 > e_s < /t_2 > \\ &\text{has the} < /t_s > [MASK] < /t_6 > \\ &\text{the event} < /t_3 > e_t < /t_4 > \end{aligned} \tag{13}$$

Given an event pair $(e_s,\ e_t)$ $(s \neq t)$ in a query sample, we define a causal label set $\gamma = \{Causality, noCausality\}$ and compute the causal score of $(e_s,\ e_t)$ based on the probability distribution over this label set at the masked position.

We construct triplet training samples $(z_q,\ z_{pos},\ z_{neg})$ consisting of the query pair and the positive and negative samples selected via the LLM. Specifically, $z_q$ represents the embedding of the original query sample after concatenation with the prompt; $z_{pos}$ refers to the embedding of the most causally plausible positive sample $p_q$, as selected by the LLM, after concatenation with the prompt; and $z_{neg}$ refers to the embedding of the least causally plausible or semantically perturbed negative sample $n_q$, also selected by the LLM and concatenated with the prompt.

In this way, prompt information is effectively incorporated into the embedding representation,

which not only enhances the models causal perception but also improves the semantic discriminability of the samples.

## D Baselines

Table 3 presents the experimental results of baseline models on the EventStoryLine and Cause-TimeBank datasets. The baselines are categorized and described as follows: **Feature-based methods: 1)Seq** (Choubey and Huang, 2017), a model for partitioning event temporal relationships in ECI. **2)DD** (Mirza and Tonelli, 2014), a data-driven method. **Knowledge-augmented methods: 1)LearnDA** (Zuo et al., 2020), a learnable knowledge-guided data augmentation method. **2)KADE** (Wu et al., 2023), a method that enhances ECI through structured knowledge and analogical reasoning. **3)DPF** (Huang et al., 2024), a method for integrating task-specific knowledge from commonsense graphs into ECI. **Graph neural network-based methods: 1)RichGCN** (Tran Phu and Nguyen, 2021), a GCN-based document-level ECI model. **2)SemSIn** (Hu et al., 2023), a semantic structure network-based method for ECI. **3)**ECLEP (Pu et al., 2023), a method that enhances ECI using event pair interaction graphs. **4)Prompt-adjusted methods: 1)KEPT** (Liu et al., 2023a), a knowledge-augmented and prompt-adjusted method for ECI. **2)DPJL** (Shen et al., 2022), a prompt-adjusted approach for enhancing ECI. **3)LKCER** (Su et al., 2025), a prompt-based approach that integrates LLM knowledge and a heterogeneous concept-level event graph for ECI. **Contrastive learning methods: 1)CLINK** (Zuo et al., 2021a), the first work introducing contrastive learning to ECI using static positive-negative pairs. **2)GCKAN** (Ding et al., 2024), a graph-based contrastive method enhanced with external knowledge. **3)ICCL** (Chao et al., 2024), an in-context contrastive learning approach for ECI.

## E Zero-shot Evaluation

To further investigate the capabilities of different model types in zero-shot ECI tasks, we compared the performance of PLMs based on masked language modeling (BERT, RoBERTa) with mainstream LLMs (GPT-3.5-turbo, GPT-4 (Goswami et al., 2020), DeepSeek-V3 (DeepSeek-AI et al., 2024)) on sentence-level ECI tasks. The experimental results are shown in Table 5. Despite LLMs having stronger global semantic understanding, their F1 scores in this task are significantly lower than those of PLMs. This phenomenon is consistent with the empirical findings of Gao et al. (Gao et al., 2023), who concluded that PLMs are better at modeling fine-grained causal relationships between events, while LLMs have certain limitations in local logical reasoning. It is worth noting that although DeepSeek-V3 performs excellently in long-text causal analysis (DeepSeek-AI et al., 2024), its sliding window attention mechanism limits its coverage of the local context, leading to poor performance in sentence-level tasks. The above results indicate a significant and close coupling between model architecture and task characteristics. MLM-driven PLMs exhibit higher adaptability for local causal identification tasks, while generative LLMs show clear disadvantages in comparison.

## F LLM-Generated Multi-Dimensional Evaluation Framework

**Prompt Input:**

You are a senior expert in causal relation evaluation. You are tasked with identifying and scoring the quality of candidate examples across the following five positive and five negative dimensions. Please generate the evaluation criteria strictly according to the procedure below:

**1. Background**

**Task:** Compare the causal relation quality between the original sentence and each candidate sentence, and produce an actionable multi-dimensional evaluation framework.

**Output Format:** A Markdown list in which each dimension includes a Name, Description, Key Points, and Example Scoring Guidelines.

**2.Positive Criteria (5 Dimensions)**

**Causal Logical Consistency:** Does the candidate sentence preserve the same causal chain as the original sentence?

**Semantic Scene Similarity:** Are the background context, theme, and participants in the candidate sentence highly aligned with those in the original sentence?

**Syntactic & Structural Fit:** Does the candidate sentence use similar connective words, syntactic patterns, and stylistic choices as the original sentence?

**Event Hierarchical Relatedness:** Do the events described in the candidate sentence correspond, at the level of concept category and verb semantics, to the core causal events in the original sentence?

**Information Completeness:** Does the candidate sentence include all essential causal elements (cause, effect, conditions, etc.) without omitting key information?

**3.Negative Criteria (5 Dimensions)**

**Causal Logical Incoherence:** Does the candidate sentence conflict with or completely invert the causal relation of the original sentence?

**Semantic Scene Disparity:** Do the theme, context, or participants in the candidate sentence diverge significantly from those in the original sentence?

**Syntactic & Structural Discrepancy:** How much does the expression style, logical connectors, and syntactic structure of the candidate sentence differ from the original?

**Event Semantic Conflict:** Do the event categories, verbs, or semantics in the candidate sentence directly conflict with those of the core event in the original sentence?

**Information Missingness or Redundancy:** Does the candidate sentence omit critical causal elements or introduce irrelevant information that disrupts overall coherence?

**4.Scoring Guidelines**

Use a 15 scale for each dimension, where 5 represents optimal performance and 1 represents the poorest performance.

For each dimension, provide one Typical Example and outline the criteria for high, medium, and low scores.

**5.Example Output Format**

**Positive Criterion 1:** Causal Logical Consistency

**Description:** The candidate sentence must maintain the same causal reasoning chain as the original sentence.

**Key Points:**

- Clear cause-then-effect ordering

- Complete and coherent reasoning chain

**Example Scoring:**

**- 5 points:** Causal relation fully consistent; reasoning is clear and natural

**- 3 points:** Generally consistent, but with occasional logical jumps or missing connectives

**- 1 point:** Causal relation is vague or completely inverted

(Subsequent dimensions follow the same format)

Please generate the full set of 5 positive and 5 negative evaluation criteria, each with corresponding example scoring guidelines, according to the requirements above.

**Final Selected LLM Output:**

**Positive Evaluation Criteria:**

**Criterion 1:** Causal Logical Consistency

**Description:** Does the candidate sentence exhibit a clear and reasonable causal relationship with the original sentence (i.e., does it reflect a similar causeeffect logic)?

**Requirements:**

Events occur in a clear cause-then-effect order

A coherent reasoning chain connects cause and effect Scoring Suggestions:

**5 points:** Fully consistent with the original causal logic; reasoning flows naturally

**3 points:** Generally consistent but somewhat forced or missing connectives

**1 point:** Causal relationship is vague or absent

**Criterion 2:** Semantic Scene Similarity

**Description:** Is the candidate sentence similar to the original in terms of semantic content or situational context?

**Requirements:**

Involves the same or similar theme (e.g., weather, behavior, social phenomena)

Shares similar subjects/objects or participant roles Scoring Suggestions:

**5 points:** Highly similar to the original; scenarios almost overlap

**3 points:** Contexts are comparable but with some differing elements

**1 point:** Themes are completely different; scenes are unrelated

**Criterion 3:** Syntactic & Structural Fit

**Description:** Does the candidate sentence maintain consistency with the original in terms of expression style and structure?

**Requirements:**

Uses similar causal connectives (e.g., because therefore, due to resulting in)

Syntax is smooth and the language style is similar Scoring Suggestions:

**5 points:** Expression style and structure are almost identical to the original

**3 points:** Basically similar with minor structural differences

**1 point:** Structure is completely different; lacks formal coherence

**Criterion 4:** Event Hierarchical Relatedness

**Description:** Do the events in the candidate sentence relate to the core causal events in the original sentence at the level of concepts, types, or verb semantics?

**Requirements:**

Events belong to the same category (e.g., natural phenomena, behavioral decisions, mental states)

Core verbs or event semantics are similar (e.g., illnesstreatment vs. injurybandaging) **Scoring Suggestions:**

**5 points:** Candidate events and original events match closely in semantic category and verb action

**3 points:** Some relatedness but at different conceptual levels

**1 point:** Events are entirely unrelated; concepts do not match

**Negative Evaluation Criteria:**

**Criterion 1:** Causal Incoherence

**Description:** Does the candidate sentence conflict with or lack coherence in its causal logic compared to the original sentence?

**Requirements:**

No clear causal reasoning chain

Events appear in reverse order or without logical precedence

May exhibit reverse causation or spurious causality

**Scoring Suggestions:**

**5 points:** Causal logic is completely contradictory or clearly wrong (e.g., effect before cause)

**3 points:** Some conflict or causal relation is ambiguous

**1 point:** No clear causality but also no direct contradiction (non-causal)

**Criterion 2:** Semantic Disparity

**Description:** Is the candidate sentence significantly different from the original in semantic theme or situational context?

**Requirements:**

Different thematic domain

Participant roles are entirely inconsistent

No overlap in context or background

**Scoring Suggestions:**

**5 points:** Semantic scenarios are completely different (e.g., weather phenomena vs. business decisions)

**3 points:** Contexts differ but remain comparably acceptable (e.g., physical discomfort vs. work stress)

**1 point:** Partial overlap in scenario, though causal logic remains unrelated

**Criterion 3:** Syntactic Mismatch

**Description:** Does the candidate sentence differ markedly from the original in structure and expression?

**Requirements:**

Lacks causal connectives (e.g., no causal conjunctions)

Grammar or expression style is inconsistent; tone is disjointed Scoring Suggestions:

**5 points:** Sentence structure is entirely different, with no causal framework

**3 points:** Noticeable structural differences, but some connectives or logical clues remain

**1 point:** Minor structural differences, overall expression still similar

**Criterion 4:** Event Conceptual Divergence

**Description:** Do the events in the candidate sentence differ or conflict significantly from the core causal events of the original in category, action, or conceptual level?

**Requirements:**

Belongs to completely different event categories (e.g., mental state vs. technical operation)

Verb semantics are in conflict or entirely unrelated

Scoring Suggestions:

**5 points:** Event types conflict or concepts are directly opposed (e.g., help vs. harm)

**3 points:** Different event categories without direct opposition

**1 point:** Different conceptual levels but with some semantic connection (e.g., cold vs. flu)

## G Prompt Design for LLM-Guided Fine-Grained Distillation and Ranking

**Prompt Input:**

You are a senior expert in causal relation analysis. Please follow the procedure below to perform fine-grained distillation and ranking of causal consistency between the main sentence and each candidate positive (or negative) example. All events are annotated with tags: <t1></t1> for the first event and <t3></t3> for the second event.

**Context Sentence:**

"sentence"

**Event 1:**

<t1>event1</t1>

**Event 2:**

<t3>event2</t3>

**Candidate Examples (Positive/Negative):**

examples_str

**Analysis & Ranking Procedure:**

**1. Semantic Parsing:**

**- Type of main event "<t1>event1</t1>":** [fill in]

**- Type of main event "<t3>event2</t3>":** [fill in]

**- Core verb:** [fill in]

**- Key participants:** [fill in]

**2. Reference Evaluation Criteria:**

- See Appendix B for positive and negative evaluation dimensions.

**3. Per-Example Scoring (Score each example from 15 on each criterion):**

Example <i>:

**Criterion 1 (Causal Logical Consistency):** [1–5] Reason: _____

**Criterion 2 (Semantic Scene Similarity/Disparity):** [1-5] Reason: _____

**Criterion 3 (Syntactic & Structural Fit/Mismatch):** [1-5] Reason: _____

**Criterion 4 (Event Hierarchical Relatedness/Conflict):** [1-5] Reason: _____

**4. Aggregate Ranking:**

**- Rank example IDs in descending order by total score:**

"[Example 3] > [Example 1] > [Example 2] > ..."

**5. Final Selection:**

**- Best Example (Positive/Negative):** [ID] [Original Example Sentence]

**- Key Evidence:** [Cite specific scoring rationale]

**- Confidence:** [Value]

**- Runner-Up Options:** [ID1, ID2]