Incorporating Temporal Coherence to Cross-Document Event Coreference Resolution

Xinyu Chen, Peifeng Li, Qiaoming Zhu*,
School of Computer Science and Technology, Soochow University, China
xychennlper@stu.suda.edu.cn, {pfli, qmzhu}@suda.edu.cn
Correspondence: qmzhu@suda.edu.cn

Abstract

Previous work on cross-document event coreference resolution (CDECR) primarily focused on enhancing semantic coherence between event mentions, largely overlooking the critical aspect of temporal coherence. To address this issue, we propose CohTP, a novel **Tem**poral **Coh**erence-driven event coreference framework. CohTP explicitly models and enforces temporal constraints by first constructing a temporal event graph via a fine-tuned natural language inference (NLI) model. The graph is then refined using an Edge-Aware GNN to resolve conflicts and partitioned into ordered time segments, where undirected edges group contemporaneous events. Event coreference resolution is subsequently performed within these temporally coherent segments, where event representations are further augmented with temporally consistent contexts. Experiments on the ECB+, GVC, WEC, and ECB+META datasets show that CohTP outperforms several state-of-the-art baselines.

1 Introduction

Cross-document event coreference resolution (CDECR) aims to cluster event mentions that refer to the same real-world occurrence across multiple text sources. This task is crucial for many downstream tasks such as question answering (Ramesh et al., 2023), topic detection (Vahidnia, 2023) and information extraction (Yan et al., 2023), While significant progress has been made in within-document settings, CDECR remains particularly challenging due to the scarcity of direct lexical and contextual cues that link mentions scattered across different documents.

Existing approaches (Cattan et al., 2021; Yu et al., 2022; Chen et al., 2025a) typically address this challenge by learning enriched representations of event mentions. As illustrated in Figure 1(a),

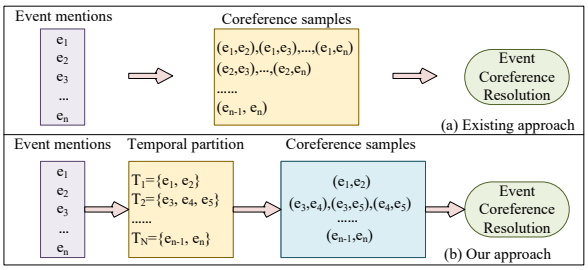

Figure 1: (a) Existing methods encode all event mention pairs and perform coreference resolution. (b) Our approach partitions it into ordered segments $\{T_1, T_2, ...\}$, and then resolves coreference only within each segment.

most previous methods overlook a fundamental temporal constraint: coreferent events must necessarily occur within overlapping or identical time periods. This oversight often leads to false positives, where semantically similar but temporally disjoint events (e.g., "the company launched product A" and "the company launched product B" months apart) are incorrectly clustered together.

To address this issue, we propose CohTP[1], a novel framework that incorporates temporal coherence into cross-document event coreference resolution. As shown in Figure 1(b), different from existing approaches, we sort the event mentions according to the time of occurrence and partition them into temporal segments for constructing candidate coreference event mention pairs in the same segments. Specifically, CohTP includes three main stages: 1) constructing and refining a temporal event graph to capture event temporal relations, 2) partitioning events into temporally coherent segments based on the refined graph, and 3) resolving coreference within each segment using temporally-augmented event representations. By restricting coreference predictions to temporally consistent segments, our framework inherently eliminates impossible coreference links and significantly reduces the candidate search space. Moreover, we enhance

*Corresponding author

[1] https://github.com/chenxinyu-nlp/CohTP

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29151–29168
July 2-7, 2026

event representations with temporally coherent contexts, selectively incorporating information from adjacent segments to incorporate temporal coherence information into event mentions for CDECR. The contributions of this paper are as follows:

- We propose CohTP, a novel temporal coherence framework that explicitly models temporal constraints for more accurate cross-document event coreference resolution.
- We introduce a temporal graph refinement approach that combines an Edge-Aware GNN with rule-based processing to ensure temporally consistent event partitions.
- We propose a temporal context augmentation method that constructs narratively coherent event representations by incorporating adjacent events, achieving performance improvements than existing coherence approaches.

2 Related Work

Event coreference resolution is a more challenging task than entity coreference resolution due to the complex structures of event mentions (Yang et al., 2015), most researchers took event coreference resolution as a pairwise similarity problem. In the within-document event coreference task, they resolved coreferent events via feature engineering (Chen and Ji, 2009; Bejan and Harabagiu, 2010; Krause et al., 2016), multi-task learning (Lu and Ng, 2017, 2021), and event representation enhancing (Tran et al., 2021; Xu et al., 2022, 2023), etc.

Early cross-document event coreference resolution task contains holistic model on nominal and verbal mentions (Lee et al., 2012), unsupervised method (Bejan and Harabagiu, 2014), iteratively unfolding inter-dependencies method (Choubey and Huang, 2017), and efficient sequential prediction paradigm(Allaway et al., 2021). Recent methodologies for CDECR can be categorized into three main strands.

Feature Representation Early approaches leveraged argument information to enrich event representations (Barhom et al., 2019; Yu et al., 2022). Subsequent work incorporated discourse structure, with Chen et al. (2023) constructing cross-document rhetorical structures and Gao et al. (2024) combining within-document rhetoric with cross-document lexical chains. Ahmed et al. (2024a) utilized cross-document Abstract Meaning Representation (X-AMR) to capture event-argument structures, which can implicitly associate temporal information via arguments like `Arg-time`. Most recently, Chen et al. (2025a) enhanced semantic coherence between event contexts through textual augmentation, demonstrating the value of discourse-level connections.

Encoder Enhancement Another line focuses on model architecture improvements. Caciularu et al. (2021) pre-trained a cross-document language model on document sets, while Held et al. (2021) developed fine-grained classifiers for local feature extraction. They aim to capture deeper linguistic patterns without explicit structural constraints.

Data Augmentation Addressing data scarcity, researchers have employed various augmentation strategies. Ahmed et al. (2023) used lemma heuristics to balance positive and negative pairs, Ravi et al. (2023) leveraged commonsense temporal relations from knowledge bases to guide event understanding, modeling generic event scripts rather than document-specific factual timelines. Ding et al. (2024) developed rationale-centric counterfactual augmentation, and Min et al. (2024) leveraged LLMs to summarize event mentions for enhanced comprehension.

Among them, Ravi et al. (2023), Ahmed et al. (2024a) and Gao et al. (2024) have integrated temporal information, which served as a supplementary feature within a unified representation or as external guidance. In contrast, our work **CohTP** is the first to: 1) explicitly model pairwise temporal relations to build a global temporal graph; 2) use the graph to partition events into temporally coherent segments; and 3) impose temporal consistency as a hard constraint by restricting coreference resolution within segments, thereby directly eliminating impossible links. This represents a paradigm shift from using time as a soft feature to leveraging it as a foundational, structural constraint.

3 Event Partition on Temporal Relation

We detail the construction and refinement of temporal graphs to partition events into ordered time segments in this section. As illustrated in Figure 2, we first employ a fine-tuned BERT-NLI model to predict temporal relations between event mention pairs and construct an initial temporal graph. The graph is then optimized through an Edge-Aware GNN with unsupervised losses to resolve temporal conflicts. Finally, events are partitioned into distinct time segments based on the refined graph.

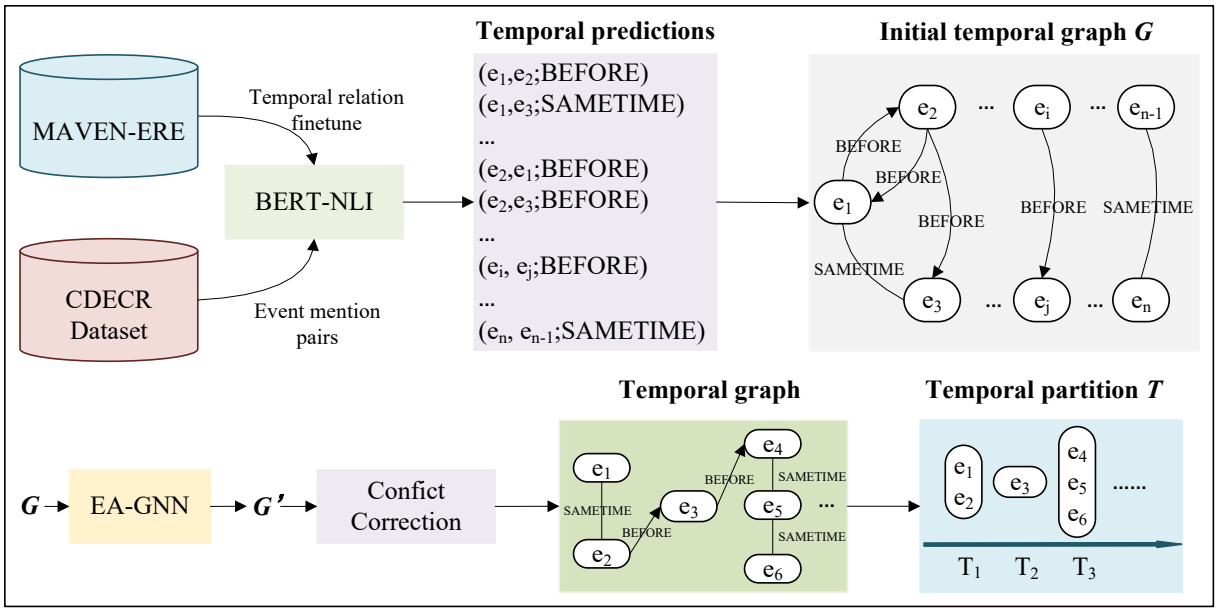

Figure 2: Framework of Temporal Graph Construction and Partitioning.

3.1 Temporal Relation Initialization

The MAVEN-ERE dataset (Wang et al., 2022) contains rich temporal annotations including BEFORE, CONTAINS, SIMULTANEOUS, OVERLAP, BEGINS-ON and ENDS-ON relations. To adapt this schema to CDECR, we consolidate CONTAINS, SIMULTANEOUS, and OVERLAP into a unified SAMETIME category, representing events occurring within the same time period. The BEFORE relation is correspondingly defined as events occurring in distinct time periods, providing simplified yet effective temporal constraints for subsequent processing.

To establish foundational temporal constraints for subsequent graph refinement, we first construct an initial temporal graph through supervised pretraining. We employ a fine-tuned Natural Language Inference (NLI) model for initial temporal relation prediction. Specifically, we adapt $\text{BERT}_{\text{Large}}$ to predict temporal relations by modeling the task as an NLI problem: for each event mention pair (e_i, e_j), we generate hypothesis templates (e.g., "Event A occurs before Event B" or "Event A occurs at the same time as Event B") and predict the temporal probability distribution $\mathbf{p}_{ij} = [p_{ij}^{\text{BF}}, p_{ij}^{\text{ST}}]$, where p_{ij}^{BF} and p_{ij}^{ST} refer to the probabilities of event mention e_i being BEFORE and SAMETIME relative to e_j respectively. To eliminate directional bias, we simultaneously predict the temporal relation of symmetric sample (e_j, e_i) to obtain the distribution $\mathbf{p}_{ji} = [p_{ji}^{\text{BF}}, p_{ji}^{\text{ST}}]$. A temporal graph $\mathcal{G}$ is initially constructed where nodes store event mention features and edges store bidirectional probability distributions. In this graph, a SAMETIME relation is represented by an undirected edge, while a BEFORE relation is represented by a directed edge pointing from the earlier event to the later one. The model achieves 78.4% accuracy on the binarized MAVEN-ERE test set.

3.2 Temporal Graph Refinement

The initial temporal graph inevitably contains prediction errors and logical inconsistencies, which require global refinement for reliable event partitioning.

Refinement Objective In the initial temporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{R})$, each node $v_i \in \mathcal{V}$ stores feature x_i of the event mention e_i, here $x_i = \text{BERT}_{\text{Large}}(\text{ctx}(e_i))$, $\text{ctx}(e_i)$ denotes the context of event mention e_i. Edges $\langle v_i, v_j \rangle \in \mathcal{E}$ store temporal relations with probability distribution $\mathbf{p}_{ij} = [p_{ij}^{\text{BF}}, p_{ij}^{\text{ST}}]$. To address the limitations of local predictions and ensure global temporal consistencies, the objective is to learn a refinement function $f_\theta(\cdot)$ that produces a conflict-minimized graph $\mathcal{G}' = f_\theta(\mathcal{G})$ that satisfies the symmetry and transitivity temporal consistency constraints:

$$\begin{aligned} p_{ij}^{\text{ST}} &= p_{ji}^{\text{ST}}, \\ p_{ij}^{\text{BF}} > \tau \wedge p_{jk}^{\text{BF}} > \tau &\Rightarrow p_{ik}^{\text{BF}} > \tau, \\ p_{ij}^{\text{ST}} > \tau \wedge p_{jk}^{\text{ST}} > \tau &\Rightarrow p_{ik}^{\text{ST}} > \tau, \end{aligned} \tag{1}$$

where τ is the confidence threshold (empirically set to 0.5) for accepting a temporal relation prediction.

Edge-Aware Graph Neural Network To effectively propagate temporal information and resolve conflicts across the graph, we propose a novel Edge-Aware Graph Neural Network (EA-GNN) to update node and edge representations.

Specifically, we first send event mention feature $\mathbf{x}_i$ to a BiLSTM to capture sequential semantics $\mathbf{h}_i^{(0)}$ and initialize the edge $\langle v_i, v_j \rangle$ representation as $\mathbf{r}_{ij}^{(0)} = \mathbf{p}_{ij}$. Second, we jointly update node and edge representations by a message passing mechanism, which incorporates edge semantics into neighbor message of j to i (denoted as $\mathbf{m}_{j\to i}^{(l)}$). We finally perform attention-based node aggregation to update node representations. The updating process can be formalized as follows.

$$\begin{aligned}
\mathbf{m}_{j\to i}^{(l)} &= \psi\left(\left[\mathbf{h}_j^{(l)} \oplus \mathbf{r}_{ij}^{(l)}\right]\right), \\
\alpha_{ij} &= \text{softmax}\left(\mathbf{a}^T \cdot \tanh(\mathbf{W}[\mathbf{h}_i^{(l)} \oplus \mathbf{m}_{j\to i}^{(l)}])\right), \\
\mathbf{h}_i^{(l+1)} &= \phi\left(\left[\mathbf{h}_i^{(l)} \oplus \sum\nolimits_j \alpha_{ij}\mathbf{m}_{j\to i}^{(l)}\right]\right), \\
\mathbf{r}_{ij}^{(l+1)} &= \phi\left(\left[\mathbf{h}_i^{(l+1)} \oplus \mathbf{h}_j^{(l+1)} \oplus \mathbf{r}_{ij}^{(l)}\right]\right),
\end{aligned} \tag{2}$$

where both $\phi(\cdot)$ and $\psi(\cdot)$ are implemented as MLP with ReLU activations, $\phi(\cdot)$ transforms edge features and $\psi(\cdot)$ processes messages for aggregation. $\oplus$ denotes concatenation, and a^T is a learnable attention vector.

After K (K is set to 3) layers, we compute the final edge probabilities as a gated combination of all layers' outputs:

$$\mathbf{r}_{ij}^{(final)} = \text{softmax}\left(\sum\nolimits_{k=1}^{K} \gamma_k \mathbf{W}^{(k)} \mathbf{r}_{ij}^{(k)}\right), \tag{3}$$

where γ_k are learned layer importance weights, and $\mathbf{W}^{(k)}$ are projection matrices. This multi-layer approach captures hierarchical temporal dependencies while mitigating gradient vanishing issues.

Temporal-aware Consistencies To further explicit supervision signals to enforce fundamental temporal logic rules, we design three unsupervised losses that directly encode temporal consistency principles at different levels.

To address common transitivity violations in the initial graph. For every event triplet (e_i, e_j, e_k), we enforce transitivity through conditional probability maximization:

$$\mathcal{L}_{\text{trans}} = -\frac{1}{|\mathcal{T}|} \sum_{(i,j,k)\in\mathcal{T}} \log\left(\mathbf{r}_{ik}^{(l)} \cdot \mathbf{v}_{ijk} + \epsilon\right), \tag{4}$$

where $\mathcal{T}$ is the set of event triplets, and $\mathbf{v}_{ijk}$ is the expected temporal vector:

$$\mathbf{v}_{ijk} = \begin{cases} [1,0]^T & \text{if } p_{ij}^{\text{BF}} > \tau \wedge p_{jk}^{\text{BF}} > \tau \\ [0,1]^T & \text{if } p_{ij}^{\text{ST}} > \tau \wedge p_{jk}^{\text{ST}} > \tau \\ \mathbf{r}_{ik}^{(0)} & \text{otherwise} \end{cases}. \tag{5}$$

To correct the prediction asymmetry caused by the directional NLI. For symmetric SAMETIME relations, we minimize distributional divergence:

$$\mathcal{L}_{\text{sym}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left\|\mathbf{p}_{ij} - \mathbf{p}_{ji}\right\|_1, \tag{6}$$

where $\mathcal{P}$ is the set of symmetric edge pairs.

To prevent over-correction, we preserve initial semantic tendencies:

$$\mathcal{L}_{\text{sem}} = \frac{1}{|\mathcal{V}|} \sum_{(i,j)\in\mathcal{V}} \text{KL}\left(\mathbf{r}_{ij}^{(l)} \oplus (g(\mathbf{h}_i^{(0)}, \mathbf{h}_j^{(0)}))\right), \tag{7}$$

where g is a temporal direction classifier pretrained on MAVEN-ERE. The combined optimization objective is:

$$\mathcal{L} = \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \lambda_{\text{sym}}\mathcal{L}_{\text{sym}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}, \tag{8}$$

with $\lambda_{\text{trans}} = 1.0$, $\lambda_{\text{sym}} = 0.6$, $\lambda_{\text{sem}} = 0.5$ empirically tuned.

These complementary losses enable holistic temporal consistency while maintaining semantic plausibility throughout the refinement process.

3.3 Training and Inference

We first fine-tune BERT-NLI[2] model on MAVEN-ERE to predict temporal relations between event mention pairs that the initial temporal probability distributions of CDECR datasets are annotated, which are used to train our EA-GNN. During inference, all samples are processed through the pretrained BERT-NLI to construct temporal graphs, which are refined by the EA-GNN for event coreference resolution.

3.4 Residual Conflict Correction

Although EA-GNN significantly reduces global temporal conflicts, its output may still contain residual inconsistencies. To ensure the reliability of subsequent segmentation steps, we apply a rule-based post-processing stage to eliminate these remaining symmetric and transitive conflicts. For symmetric conflicts, if an edge (e_i, e_j) is predicted as

[2]transformers.BertForSequenceClassification

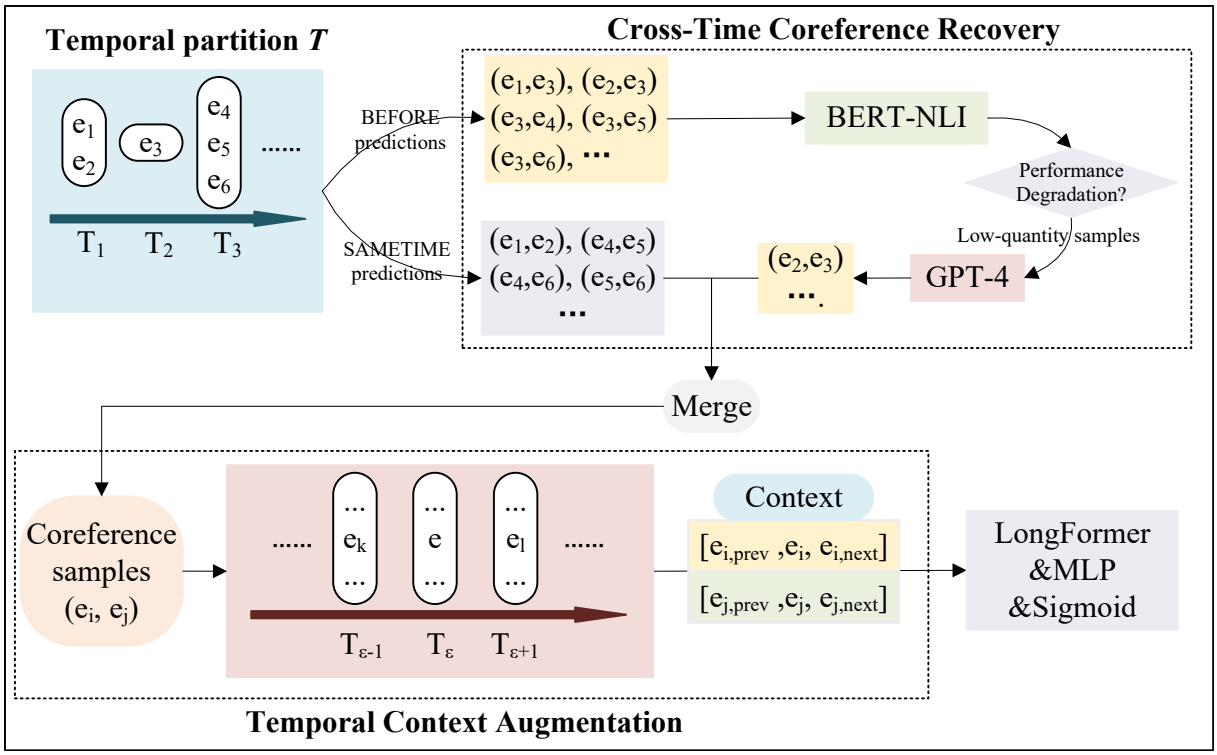

Figure 3: Framework of Temporal-aware Coreference Resolution.

SAMETIME while its symmetric edge (e_j, e_i) is BEFORE, we adopt the relation with higher confidence, that is, we remove the one with lower confidence and retain the one with higher confidence. For transitive conflicts[3], we examine all event mention triplets (e_i, e_j, e_k). When an inconsistency is detected, we calculate the joint probability sum of all possible temporal relation combinations and correct the predictions of the edges within the triplet to the conflict-free combination with the highest joint probability. This step ensures the logical consistency of the temporal graph.

3.5 Temporal Partitioning

We partition the event mentions into ordered time segments $\mathcal{T} = \{T_1, T_2, ..., T_M\}$ based on the obtained temporal graph. This partition satisfies: 1) events connected by undirected SAMETIME edges are grouped into the same segment T_m, forming clusters of contemporaneous events; 2) and the directed 'BEFORE' edges establish the temporal ordering between segments, ensuring that for any $a < b$, all events in T_a temporally precede those in T_b. This structure enables us to represent complex temporal relationships through a simplified timeline of event clusters.

[3]Details and an illustrative example are available in Appendix A.

4 CDECR with Temporal Coherence

Our framework of temporal coherence-enhanced approach for cross-document event coreference resolution is shown in Figure 3. Following temporal partitioning, we first employ a cross-time coreference recovery mechanism that utilizes large language models to correct low-quality temporal predictions between adjacent segments. For candidate coreferent event mention pairs within the same time segment, we then construct temporally coherent contextual representations, which are encoded by LongFormer and processed through an MLP for coreference prediction.

4.1 Cross-Time Coreference Recovery

To alleviate the risk of incorrectly separating coreferent events across different time segments due to faulty temporal predictions, we correct potentially erroneous BEFORE predictions between adjacent segments by a cross-time coreference recovery mechanism. Specifically, we collect event pairs predicted as BEFORE between neighboring segments $T_{\varepsilon\pm1}$ and the target segment T_ε, and fine-tune the BERT-NLI model on these samples while monitoring performance on the MAVEN-ERE temporal test set (detailed in Appendix B). Samples whose inclusion causes performance degradation are regarded as low-quality predictions, which account

for approximately 32% of all BEFORE predictions. These samples are subsequently submitted to GPT-4 for re-evaluation[4], with approximately 65% of them being corrected (from BEFORE to SAME-TIME). This intervention effectively recovers coreferent events that were incorrectly dispersed across different segments due to initial temporal prediction errors. We also report the performance using other LLMs for temporal correction in Appendix C.

4.2 Temporal Context Augmentation

Following temporal partitioning, we perform coreference resolution enhanced by temporal coherence. For candidate coreferent event pairs (e_i, e_j) within the same time segment T_ε, we augment their contextual representations to improve disambiguation. Specifically, we employ the coherence computation module $Coh(\cdot)$ proposed by Chen et al. (2025a) to select the most coherent contextual events e_{prev} from the preceding segment $T_{\varepsilon-1}$ and e_{next} from the subsequent segment $T_{\varepsilon+1}$ for both e_i and e_j:

$$\begin{aligned} e_{i,\text{prev}} &= \text{argmax}_{e_k \in T_{\varepsilon-1}} \text{Coh}([\mathbf{h}_k, \mathbf{h}_i]), \\ e_{i,\text{next}} &= \text{argmax}_{e_l \in T_{\varepsilon+1}} \text{Coh}([\mathbf{h}_i, \mathbf{h}_l]). \end{aligned} \tag{9}$$

We then construct augmented sequences $\mathcal{S}_i$ and $\mathcal{S}_j$ for event mention e_i and e_j as follows.

$$\begin{aligned} \mathcal{S}_i &= [\text{ctx}(e_{i,\text{prev}}); \text{ctx}(e_i); \text{ctx}(e_{i,\text{next}})], \\ \mathcal{S}_j &= [\text{ctx}(e_{j,\text{prev}}); \text{ctx}(e_j); \text{ctx}(e_{j,\text{next}})], \end{aligned} \tag{10}$$

where $\text{ctx}(e)$ is the context of event mention e.

4.3 Coreference Scoring

We encode $\mathcal{S}_i$ and $\mathcal{S}_j$ by LongFormer (Beltagy et al., 2020) to obtain representations $\mathbf{f}_i$ and $\mathbf{f}_j$. Then we compute the coreference confidence score through a multi-layer perceptron (MLP):

$$\theta_{ij} = \text{MLP}([\mathbf{f}_i; \mathbf{f}_j]), S_{ij} = \text{Sigmoid}(\theta). \tag{11}$$

4.4 Training Objective and Inference

We train our coreference model by minimizing the binary cross-entropy loss over all candidate event mention pairs. Given the set of all candidate pairs $\mathcal{P}$ and their coreference labels $y_{ij} \in \{0, 1\}$, the loss function is defined as:

$$\begin{aligned} \mathcal{L}_{\text{cr}} = -\frac{1}{|\mathcal{P}|} \sum\nolimits_{(i,j)\in\mathcal{P}} [y_{ij} \log S_{ij} + \\ (1 - y_{ij}) \log(1 - S_{ij})]. \end{aligned} \tag{12}$$

During inference, we group the test set documents using the method of Barhom et al. (2019). Within each resulting topic cluster, we treat all event mention pairs as candidates for coreference. These candidate pairs are then processed by our CDECR model to generate pairwise coreference scores. Finally, we apply best-first clustering (Huang et al., 2019) to these scored pairs to form the final coreference chains.

5 Experiment and Discussion

5.1 Experimental Settings

Datasets We evaluate our event coreference model on ECB+ (Cybulska and Vossen, 2014), GVC (Vossen et al.), WEC (Eirew et al., 2021) and ECB+META (Ahmed et al., 2024b) datasets. To ensure comparability with prior work at the foundational level, we aligns with established benchmarks in two key aspects: 1) using gold-standard data for both training and inference, 2) and maintaining consistency of data partitioning with Bugert et al. (2021), and Ahmed et al. (2024b) on ECB+, GVC, and ECB+META datasets respectively. In addition, we provide a separate comparison with Cattan et al. (2021) under the setting of extracted mentions in Appendix D.

Metrics and Hyperparameters Pairwise F1 and the overall CoNLL-F1 score is used to streamline evaluation and enhance interpretability. The details of metrics are available in Appendix E. Besides, the details of hyperparameters and the performance under different configurations are available in Appendix G.

Backbones We evaluate CohTP with different encoder combinations, which is available in Appendix F.

5.2 Experimental Results

Table 1 shows evaluation results over the ECB+, GVC, WEC, ECB+META1 and ECB+METAm test sets. Our results include comparisons with high-performing systems from recent years and several different aspect of ablations. The details of existing baselines are available in Appendix H.

[4]We randomly select 50 samples of temporal relation extraction tasks in MAVEN-ERE, and used ChatGPT, GPT-4 and LLaMa-7B to predict the temporal relations of these 50 samples respectively, and achieved accuracy of 54%, 62% and 60% respectively, so we choose GPT-4 to intervene in low-quality prediction. The prompt of temporal relation re-evaluation is available in Appendix L.1.

	System	ECB+	GVC	WEC	ECB+META1	ECB+METAm
Existing Baselines	CD-DCE	76.9 | 88.5	- | 87.3	58.8 | 65.9	61.4 | 72.0	46.1 | 56.2
	MP	- | -	- | -	- | -	- | 71.4	- | 55.6
	KD	- | 86.4	- | 83.0	- | -	- | -	- | -
	LLM-Min	- | 86.7	- | 87.4	- | -	- | -	- | -
	DSSI	- | 85.5	- | -	- | 65.0	- | -	- | -
	CD-DRS	- | 86.4	- | -	- | -	- | 69.2	- | 51.8
	HT	67.7 | 77.2	68.1 | 73.4	40.3 | 48.5	46.1 | 56.2	42.3 | 50.1
LLM Intervention	All Samples	74.3 | 85.2	73.6 | 80.8	55.3 | 61.4	56.3 | 65.1	41.7 | 49.6
	No Intervention	68.5 | 79.3	64.8 | 70.6	56.8 | 62.5	59.4 | 58.2	38.3 | 44.1
LLM Substitutability	LLaMa-Only	63.1 | 71.4	58.8 | 65.7	49.7 | 55.4	53.4 | 62.2	41.1 | 47.5
	LLaMa+Temp	65.8 | 76.8	63.1 | 71.2	51.8 | 58.0	58.1 | 66.0	42.9 | 49.7
Our Model	CohTP	**77.4 | 88.9**	**80.4 | 87.7**	**59.7 | 66.7**	**62.8 | 72.2**	**48.3 | 56.6**
Temporal Impact	w/o Temporal	76.4 | 86.6	78.9 | 85.8	58.5 | 65.1	61.6 | 70.1	46.7 | 54.3
	w/o Temp&Coh	74.9 | 83.9	76.6 | 80.4	56.6 | 63.6	59.5 | 68.4	44.2 | 50.5
Graph Refinement	w/o Trans	76.6 | 87.4	78.7 | 85.5	56.0 | 63.1	62.3 | 70.4	44.5 | 52.9
	w/o Sym	76.8 | 88.1	79.0 | 86.1	57.9 | 64.7	62.0 | 71.0	45.4 | 53.3
	w/o Sem	77.1 | 88.3	79.5 | 86.2	58.2 | 64.9	62.6 | 70.9	45.8 | 53.6

Table 1: Pairwise | CoNLL-F1 scores of cross-document event coreference resolution on ECB+, GVC, WEC, ECB+META1 and ECB+METAm datasets, where "-" represents the result that has not been reported. The complete metrics including MUC, B^3, and $CEAF_e$ for all datasets are provided in Appendix J Table 12, 13, 14, 15, 16.

Our method CohTP significantly ($P < 0.01$, t-test) outperforms most of these baselines on CoNLL-F1 score, which indicates the effectiveness of temporal coherence in CDECR.

Compared with the SOTA baseline CD-DCE, which relies on computationally expensive discourse rhetoric structure (DRS), our CohTP offers a more efficient pathway by leveraging temporal coherence. While CD-DCE's complex structural modeling achieves strong performance, our method achieves comparable (even slightly superior on ECB+, with +0.5 Pairwise-F1 and +0.4 CoNLL-F1) results without the need for complex discourse parsing. This demonstrates that temporal constraints provide a computationally more efficient and equally effective inductive bias for event coreference resolution. On average across all evaluated datasets, our CohTP processed each event mention pair in 1.1 seconds, significantly faster than CD-DCE's 2.1 seconds. CohTP is nearly twice as fast, demonstrating its superior overall efficiency. A detailed breakdown is provided in Appendix I.

5.3 Ablation and Discussion

Impact of Temporal Information In Table 1, "w/o Temporal" represents using only the coherence computation function $\text{Coh}(\cdot)$ to match texts that are semantically coherent with the specific event mention context, thereby obtaining enhanced narrative texts [$\text{ctx}(e_{\text{prev}})$, $\text{ctx}(e)$, $\text{ctx}(e_{\text{next}})$] for event coreference resolution, with the entire process excluding any temporal information (e.g., temporal graphs, temporal partitioning), "w/o Temp&Coh" represents removing coherence information from the "w/o Temporal" setup, i.e., using only $\text{ctx}(e)$ for event coreference resolution. It can be seen that removing temporal cues (w/o Temporal) reduces CoNLL-F1 by 2.3 points on ECB+ (88.9→86.6) and 1.9 points on GVC (87.7→85.8). When both temporal and coherence information are eliminated (w/o Temp & Coh), the performance drop widens to 5.0 points on ECB+ (88.9→83.9) and 7.3 points on GVC (87.7→80.4). This substantial degradation demonstrates that while coherence alone aids in event coreference resolution, the incorporation of temporal information is crucial for imposing hard constraints against impossible coreference links across different time periods, thereby effectively reducing false positives and leading to more robust coreference resolution.

Impact of LLM Intervention Strategies GPT-4 intervening on all samples (All Samples) lowers ECB+ CoNLL-F1 by 3.7 points (88.9 →85.2), as blanket corrections introduce noise. Without any intervention (No Intervention), performance drops sharply by 9.6 points (88.9→79.3), resulting in more false coreference due to unaddressed SAME-TIME errors. In contrast, CohTP corrects about 65% of critical erroneous BEFORE predictions, boosting recall by 2.8% without sacrificing precision. It indicates that CohTP achieves an optimal balance between precision and recall.

Besides, the analysis on LLM Substitutability and Graph Refinement impact are available in Appendix J.3 and J.5.

Recover Window Analysis Additionally, we also extend the cross-time recovery mechanism beyond adjacent segments to assess its potential for recovering coreferent pairs separated by larger temporal gaps. Specifically, we varied the recovery window δ in the target segment T_ε, considering $T_{\varepsilon\pm\delta}$ for δ = 0 (no intervention), 1 (adjacent only, our original setting), 2, 3, and 4. Table 2 reports the results on ECB+.

RW	RR(%)	P	R	F1	CoNLL
δ=0	0.0	72.8	64.7	68.5	79.3
$\delta\leq$1	64.9	**90.7**	**67.5**	**77.4**	**88.9**
$\delta\leq$2	69.5	89.5	68.0	77.3	88.5
$\delta\leq$3	71.1	88.6	68.4	77.2	88.3
$\delta\leq$4	72.0	86.8	68.7	76.7	88.1

Table 2: Results on different sizes of recovery window. RW=Recovery Window, RR=Recovery Rate. P, R, and F1 represent the scores of precision, recall, and F1 of pairwise prediction, respectively.

Our original design (δ=1) strikes the optimal balance, achieving the highest F1 scores. Extending recovery to non-adjacent segments introduces more false positives (hurting precision) and offers only marginal recall gains, leading to no improvement (or even degradation) in overall coreference performance. Given that only 3.5% of gold coreferent pairs are separated by $\geq$2 segments (as shown in our multi-segment split analysis), the practical benefit of broader recovery is limited. Nevertheless, we appreciate this suggestion and will include this analysis in the final version, while noting that exploring more sophisticated multi-hop recovery methods (e.g., graph-based propagation) is a promising direction for future work.

5.4 Error Analysis

Errors of Temporal Relation Prediction To directly quantify the effect of our temporal relation prediction, we conducted an analysis on a randomly selected topic from the ECB+ dataset. 27.8% of the gold-standard coreferent pairs were incorrectly predicted as BEFORE, these pairs were placed into separate temporal segments and excluded from coreference resolution, constituting the primary source of recall loss introduced by our framework. Among the remaining candidate coreferent pairs, compared to the baseline without any temporal information, the inclusion of it corrected erroneous coreference links for 68.1% of these pairs, while maintaining originally correct predictions for 22.2%, while only 9.7% introduce new errors.

These error and correction cases reveal the layered nature of temporal information expression in text. Cases misjudged as BEFORE often rely heavily on discourse-level inference or external knowledge, lacking explicit temporal adverbials on the surface syntactic level, which makes it difficult for the model to capture their sequential relationship. In contrast, the successfully corrected SAMETIME cases involve distinct event arguments (e.g., different participants or locations), where semantic models tend to conflate them while the temporal model can effectively differentiate. The minority of cases that introduce new errors mostly involve ambiguities in temporal expression.

Errors in Graph Refinement We analyze temporal consistency in refined graphs from the ECB+ test set, where refinement quality is measured by conflict rate reduction (we set 60% as the threshold) after processing. In well-refined subgraphs with over 60% conflict reduction, 73.3% of cases correct coreference errors present in the initial graph, 14.6% maintain originally correct predictions, while 12.1% introduce new inconsistencies. Mechanistic analysis indicates that successful refinement typically occurs in densely-connected subgraphs where temporal information can propagate effectively through multiple paths, allowing the EA-GNN to resolve local prediction errors via global consistency constraints.

This substantial error correction rate demonstrates that effective subgraph refinement significantly improves coreference accuracy. For poorly-refined subgraphs with under 60% conflict reduction, 61.4% maintain original results, 29.8% introduce additional coreference errors, and only 8.8% correct existing mistakes. We observe that these subgraphs often suffer from sparse connectivity or contain clusters of mutually reinforcing incorrect predictions that create stable but erroneous local minima during refinement. This pattern indicates that inadequate refinement of subgraphs primarily degrades rather than enhances coreference resolution performance, particularly when the initial temporal predictions contain systematic biases in specific event types.

Impact of LLM Intervention Examining all LLM-corrected temporal relations reveals that 77.2%

of interventions properly resolve coreference errors, 12.9% maintain previous correct decisions, while 9.9% incorrectly modify valid temporal relations. The high success rate (77.2%) demonstrates LLM's effectiveness in handling challenging temporal cases, particularly in scenarios requiring commonsense understanding of event durations and typical sequences that may not be explicitly stated in text. The 9.9% error rate suggests need for more precise intervention criteria, with analysis showing that these errors often occur when documents contain technical or domain-specific temporal expressions that fall outside the LLM's training distribution. Notably, 80.5% of successful interventions involve correcting BEFORE relations to SAME-TIME, confirming that cross-time coreference recovery is the primary benefit of LLM integration. This pattern aligns with our expectation that LLMs excel at identifying contemporaneous events described with varying temporal expressions, whereas our base temporal model might over-segment these into separate time periods due to surface-level linguistic differences.

5.5 Case Study

We present a concrete case from the ECB+ test set to illustrate how our temporal coherence framework resolves coreferent event mentions. The example involves multiple earthquake type events across eight sentences.

Context S1: Indonesia 's West Papua province was hit by a magnitude 6.1 ⟨earthquake⟩ today , the latest powerful tremor to shake the region where five people were killed and hundreds injured at the weekend when buildings were destroyed .

Context S2: A strong ⟨earthquake⟩ rattled Indonesia 's West Papua province Wednesday just days after a powerful quake levelled buildings and killed one person .

Context S3: Atururi said a 10-year-old girl was ⟨killed⟩ and at least 40 people were injured in the ⟨earthquakes⟩ , which rekindled bitter memories of similar deadly quakes that hit the town in 2002 .

Context S4: A series of powerful ⟨earthquakes⟩ rocked Manokwari , the capital of West Papua , on Sunday , killing four people , injuring dozens and destroying hundreds of buildings .

Context S5: A series of earthquakes ⟨killed⟩ a 10-year-old girl and injured dozens Sunday in remote eastern Indonesia and briefly triggered fears of another tsunami in a country still recovering from such a disaster in 2004 .

Consider three key event mentions: the ⟨earthquake⟩ in context S1 ("Indonesia's West Papua province was hit by a magnitude 6.1 earthquake today"), the ⟨earthquake⟩ in context S2 ("A strong earthquake rattled Indonesia's West Papua province Wednesday"), and the ⟨earthquakes⟩ in context S4 ("A series of powerful earthquakes rocked Manokwari... on Sunday"). While these mentions share similar semantics and location, they refer to distinct seismic events occurring at different times.

Our temporal analysis successfully distinguishes these events through explicit time cues: "today" in S1, "Wednesday" in S2, and "Sunday" in S4. The BERT-NLI classifier identifies these as temporally distinct occurrences, while the EA-GNN refinement reinforces these temporal distinctions in the graph structure. The temporal partitioning module subsequently places these events into separate time segments, preventing them from being considered as coreference candidates.

Meanwhile, our method correctly identifies coreferent events that share both semantic and temporal consistency. For instance, the ⟨killed⟩ events in context S3 ("a 10-year-old girl was killed") and context S5 ("A series of earthquakes killed a 10-year-old girl") are properly clustered together as they refer to the same casualty incident within compatible timeframes.

This case demonstrates two key advantages of our approach: it effectively separates temporally distinct but semantically similar events that would cause false positives in baseline methods, while successfully identifying genuine coreference links through joint temporal-semantic analysis. The explicit modeling of temporal constraints proves particularly valuable in news scenarios where multiple related events occur in close succession but represent distinct real-world occurrences.

6 Conclusion

We proposed a novel framework CohTP for CDECR that leverages the constraint that coreferent events must occur within overlapping time periods. Our approach constructs and refines temporal graphs to partition events into coherent segments, then performs coreference resolution within these segments while incorporating temporally-aware context augmentation. Experiment results on several datasets demonstrate that our proposed method outperforms several state-of-the-art baselines.

Limitations

Our method still suffers from several shortcomings, which will be addressed in our future work. First, we only perform coreference resolution on golden mentions. The upstream task span detection is also important for coreference resolution. Second, the performance of our temporal segmentation framework heavily relies on the initial temporal relation predictions. Errors in the temporal graph construction phase can propagate to subsequent coreference resolution. Third, the granularity of temporal partitioning presents challenges, overly fine-grained segments may separate truly coreferent events, while overly coarse segments may include temporally incompatible events. Fourth, in the rule-based post-processing for resolving transitive conflicts, we employ a sum of probabilities as a practical heuristic to approximate joint likelihood. While effective (as shown in ablation studies), a product of probabilities would be more principled from a probabilistic graphical model perspective. Exploring this represents a valuable direction for future refinement of the framework. In future work, we plan to develop more robust temporal relation classifiers that better handle complex event durations and fuzzy temporal expressions. We will also explore adaptive temporal partitioning strategies that dynamically adjust segment granularity based on event density and characteristics. Finally, we aim to extend our temporal coherence framework to multilingual and multimodal settings, exploring methods to establish temporal consistency across different languages and modalities (e.g., text and video), thereby enhancing the broader applicability of our approach.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62376181 and 62276177), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Shafiuddin Rehan Ahmed, George Arthur Baker, Evi Judge, Michael Regan, Kristin Wright-Bettner, Martha Palmer, and James H. Martin. 2024a. Linear cross-document event coreference resolution with X-AMR. In *LREC/COLING*, pages 10517–10529. ELRA and ICCL.

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. 2023. 2*n is better than n^2: Decomposing event coreference resolution into two tractable problems. In *ACL (Findings)*, pages 1569–1583.

Shafiuddin Rehan Ahmed, Zhiyong Eric Wang, George Baker, Kevin Stowe, and James H. Martin. 2024b. Generating harder cross-document event coreference resolution datasets using metaphoric paraphrasing. In *ACL (Short Papers)*, pages 276–286.

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. In *EMNLP (1)*, pages 4659–4671. Association for Computational Linguistics.

Amit Bagga. 1998. Evaluation of coreferences and coreference resolution systems. In *LREC*, pages 563–572.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *ACL*, pages 4179–4189.

Cosmin Adrian Bejan and Sanda M. Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *ACL*, pages 1412–1422.

Cosmin Adrian Bejan and Sanda M. Harabagiu. 2014. Unsupervised event coreference resolution. *Comput. Linguistics*, 40(2):311–347.

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

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Comput. Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: cross-document language modeling. In *EMNLP (Findings)*, pages 2648–2662.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. In *ACL/IJCNLP (Findings)*, pages 5100–5107.

Xinyu Chen, Peifeng Li, and Qiaoming Zhu. 2025a. Employing discourse coherence enhancement to improve cross-document event and entity coreference resolution. In *ACL (1)*, pages 23272–23286. Association for Computational Linguistics.

Xinyu Chen, Peifeng Li, and Qiaoming Zhu. 2025b. Improving cross-document event coreference resolution by discourse coherence and structure. *Inf. Process. Manag.*, 62(4):104085.

Xinyu Chen, Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Cross-document event coreference resolution on discourse structure. In *EMNLP*, pages 4833–4843.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *ACL*, pages 54–57.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *EMNLP*, pages 2124–2133.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*, pages 4545–4552.

Bowen Ding, Qingkai Min, Shengkun Ma, Yingjie Li, Linyi Yang, and Yue Zhang. 2024. A rationale-centric counterfactual data augmentation method for cross-document event coreference resolution. *CoRR*, abs/2404.01921.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: deriving a large-scale cross-document event coreference dataset from wikipedia. In *NAACL-HLT*, pages 2498–2510.

Qiang Gao, Bobo Li, Zixiang Meng, Yunlong Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024. Enhancing cross-document event coreference resolution by discourse structure and semantic information. In *LREC-COLING*, pages 5907–5921.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *EMNLP*, pages 1406–1417.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. 2019. Improving event coreference resolution by learning argument compatibility from unlabeled data. In *NAACL-HLT*, pages 785–795.

Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn. 2016. Event linking with sentential features from convolutional neural networks. In *CoNLL*, pages 239–249.

Heeyoung Lee, Marta Recasens, Angel X. Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *EMNLP-CoNLL*, pages 489–500.

Jing Lu and Vincent Ng. 2017. Joint learning for event coreference resolution. In *ACL*, pages 90–101.

Jing Lu and Vincent Ng. 2021. Span-based event coreference resolution. In *AAAI*, pages 13489–13497.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT-EMNLP*, pages 25–32.

Qingkai Min, Qipeng Guo, Xiangkun Hu, Songfang Huang, Zheng Zhang, and Yue Zhang. 2024. Synergetic event understanding: A collaborative approach to cross-document event coreference resolution with large language models. In *ACL*, pages 2985–3002.

Abhijnan Nath, Shadi Manafi, Avyakta Chelle, and Nikhil Krishnaswamy. 2024. Okay, let's do this! modeling event coreference with generated rationales and knowledge distillation. *CoRR*, abs/2404.03196.

Gowtham Ramesh, Makesh Narsimhan Sreedhar, and Junjie Hu. 2023. Single sequence prediction over reasoning graphs for multi-hop QA. In *ACL*, pages 11466–11481.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. In *EACL*, pages 1700–1716.

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In *ACL-IJCNLP*, pages 4840–4850.

Sahand Vahidnia. 2023. *Deep and Temporal Ontology Guided Clustering Methods and Representation Learning for Topic Detection and Tracking*. Ph.D. thesis.

Marc B. Vilain, John D. Burger, John S. Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC*, pages 45–52.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. Don't annotate, but validate: a data-to-text method for capturing event data. In *LREC*, pages 3034–3042.

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 *EMNLP*, pages 926–941. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *ACL (1)*, pages 2526–2547. Association for Computational Linguistics.

Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2022. Improving event coreference resolution using document-level and topic-level information. In *EMNLP*, pages 6765–6775.

Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Coref-prompt: Prompt-based event coreference resolution by measuring event type and argument compatibilities. In *EMNLP*, pages 15440–15452.

Hang Yan, Yu Sun, Xiaonan Li, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. 2023. UTC-IE: A unified token-pair classification architecture for information extraction. In *ACL*, pages 4096–4122.

Bishan Yang, Claire Cardie, and Peter I. Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *Trans. Assoc. Comput. Linguistics*, 3:517–528.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022. Pairwise representation learning for event coreference. In **SEM@NAACL-HLT*, pages 69–78.

A Rule-based Residual Conflict Correction

A.1 Conflict Detection

We identify two primary types of temporal conflicts in the refined temporal graph:

- Symmetric Conflicts: Occur when (e_i, e_j) is predicted as SAMETIME but (e_j, e_i) is predicted as BEFORE, violating the symmetry property.
- Transitive Conflicts: Arise from violations of temporal transitivity rules across event triplets (e_i, e_j, e_k), where the relations between (e_i, e_j) and (e_j, e_k) imply a specific relation between (e_i, e_k) that contradicts the model's prediction.

Table 3 shows the transitive conflict patterns for event triplet (e_i, e_j, e_k).

Pattern	(e_i, e_j)	(e_j, e_k)	(e_i, e_k)	Conflict
1	B	B	B	
2	B	B	S	✓
3	B	S	B	
4	S	B	B	✓
5	S	B	B	
6	S	B	S	✓
7	S	S	B	✓
8	S	S	S	

Table 3: Transitive Conflict Patterns for Event Triplet (e_i, e_j, e_k). B = BEFORE, S = SAMETIME; ✓ indicates conflict presence.

A.2 Conflict Correction

For symmetric conflicts, we adopt the relation with higher confidence score, removing contradictory bidirectional connections and retaining only the unidirectional relation.

For transitive conflicts, we employ a joint probability maximization approach. Given initial predictions:

- $\text{pred}(e_i, e_j) = (\text{SAME}, [0.2, 0.8])$,
- $\text{pred}(e_j, e_k) = (\text{BEFORE}, [0.9, 0.1])$,
- $\text{pred}(e_i, e_k) = (\text{SAME}, [0.3, 0.7])$,

we calculate joint probabilities for valid patterns as table 4:

Pattern	(e_i, e_j)	(e_j, e_k)	(e_i, e_k)	Sum
B-B-B	0.2	0.9	0.3	1.4
B-S-B	0.2	0.1	0.3	0.6
S-B-B	0.8	0.9	0.3	2.0
S-S-S	0.8	0.1	0.7	1.6

Table 4: Probability calculation example.

Dataset	Sum	Product	Δ(Sum - Product)
ECB+	88.9	88.7	+0.2
GVC	87.7	87.4	+0.3
WEC	66.7	66.1	+0.6
ECB+META1	72.2	71.7	+0.5
ECB+METAm	56.6	56.2	+0.4

Table 5: CoNLL-F1 scores for different conflict resolution strategies.

The S-B-B pattern achieves the highest joint probability (2.0), therefore we correct the predictions to:

- (e_i, e_j): SAMETIME
- (e_j, e_k): BEFORE
- (e_i, e_k): BEFORE (corrected from SAMETIME).

A.3 Comparison of Conflict Resolution Strategies

We conducted a comparative experiment replacing the heuristic sum of probabilities with the more principled product of probabilities. Table 5 shows the performance difference on the CoNLL-F1 metric across all datasets.

While the product of probabilities is theoretically more grounded for assessing joint likelihood, the heuristic sum of probabilities yields

marginally superior performance across all benchmarks. The consistent but small performance gap (0.2-0.6 F1 points) suggests that in the context of our temporally-constrained graph, the sum heuristic serves as a robust and effective approximation for selecting the most consistent temporal pattern. Therefore, we retain the sum of probabilities in our final model, as it provides better performance.

B Details of Performance Monitoring

To identify candidate samples whose labels are likely incorrect and would degrade the model's temporal reasoning capability if learned. We monitor a highly sensitive metric—the aggregate shift in the model's predictive confidence for the correct labels across the entire MAVEN-ERE test set. This detects harmful "cognitive drift" even when no prediction flips from correct to incorrect.

For each candidate sample S, we perform a minimal update (a few steps with a low learning rate) and compute the sum of probability changes for the correct class over all test samples. Consider a test set with 4 samples. Table 7 shows the model's predictions before and after the update with the candidate S.

The Total Aggregate Impact is: (+0.0002) + (-0.0001) + (+0.0003) + (-0.0006) = -0.0002.

A negative total proves that S's influence, in aggregate, degrades the model's confidence on established facts, thus flagging it as a low-quality sample. This method is far more sensitive and principled than monitoring accuracy.

This diagnostic is computationally manageable because it is highly parallelizable. Each sample's test is independent. On an RTX 4090, we can run 6 diagnostics concurrently. This reduces the total wall-clock time for processing all low-confidence samples to a highly practical about 2 hours. This targeted investment is justified, as it precisely identifies the 32% of samples that benefit most from LLM correction, leading to significant overall performance gains.

C Temporal Correction by Other LLMs

Table 6 presents the zero-shot and few-shot temporal correction performance of GPT-3.5-Turbo, LLaMa-7B, Flan-T5, and our GPT-4 model across all evaluation datasets.

D Performance on Extracted Mentions

Following established practices (e.g., Cattan et al. (2021)), we evaluated our method using both gold and predicted mentions in Table 8 and the results show the CoNLL-F1 is 62.8, much lower than that using gold mentions (88.9). The performance drop (-26.1) is consistent with patterns in prior work (e.g., Cattan et al. (2021) reported a drop from 81.0 to 54.4), highlighting the challenge of error propagation from mention detection. Despite this, our temporal coherence framework provides a relative advantage, and we plan to explore joint modeling for mention detection and coreference in future work.

E Metrics

While previous studies often report MUC (Vilain et al., 1995), B^3 (Bagga, 1998), and $CEAF_e$ (Luo, 2005) metrics individually, we focus on Pairwise F1 and the overall CoNLL-F1 score to streamline evaluation and enhance interpretability. The Pairwise F1 score offers a direct assessment of the model's ability to perform the fundamental task of classifying coreference links, which is the basis of our approach. The CoNLL-F1 score, as the standard blend of MUC, B^3 , and $CEAF_e$, provides a unified and robust summary of system performance, mitigating the bias of any single metric and facilitating direct comparison with the majority of existing literature.

F Backbones Evaluation

We experiment with DeBERTa-v3-large(He et al., 2021) and ModernBERT-large(Warner et al., 2025) for both temporal relation prediction and coreference resolution. Table 10 reports the CoNLL-F1 scores on ECB+.

G Hyperparameters

For temporal relation initialization, we fine-tune BERT-large with a batch size of 16 over 5 epochs, using a learning rate of 2e-5 and dropout rate of 0.1. The graph layers and hidden states dimensional of EA-GNN are set to 3 and 256 respectively. We train our EA-GNN with a batch size of 8 for 50 epochs at learning rate 5e-5 and dropout 0.2. In the temporal-aware coreference resolution stage, we use Longformer-base to encode augmented event sequences, with batch size 8, 10 training epochs, learning rate 1e-5, and dropout 0.1. Our imple-

LLM	Setting	ECB+	GVC	WEC	ECB+META1	ECB+METAm
No Invervention	-	79.3	70.6	62.5	58.2	44.1
GPT-3.5-Turbo	Zero-shot	84.7	76.5	64.6	65.0	48.5
	Few-shot(k=4)	86.0	78.8	65.8	67.5	50.8
LLaMa-7B	Zero-shot	81.5	73.9	63.0	62.0	46.0
	Few-shot(k=4)	83.2	76.0	64.5	64.8	48.5
Flan-T5	Zero-shot	80.8	73.2	62.8	61.5	45.5
	Few-shot(k=4)	82.5	75.4	64.2	64.0	47.9
GPT-4 (Ours)	Zero-shot	87.5	83.0	65.5	70.5	54.0
	Few-shot(k=4)	**88.9**	**87.7**	**66.7**	**72.2**	**56.6**

Table 6: CoNLL-F1 of Large Language Models for Temporal Correction

TS	$\text{Pred}_{\text{orig}}$	$\text{Pred}^{\text{S}}_{\text{update}}$	IV
s1	[0.3, 0.7];(ST)	[0.298, 0.702]	+0.002 (P)
s2	[0.8, 0.2];(ST)	[0.801, 0.199]	-0.001 (N)
s3	[0.6, 0.4];(BF)	[0.603, 0.397]	+0.003 (P)
s4	[0.1, 0.9];(BF)	[0.094, 0.906]	-0.006 (N)

Table 7: Predictions before and after the update. TS=Test Sample, $\text{Pred}_{\text{orig}}$ represents original prediction, whose format is "$[p_{BF}, p_{ST}]$; (gold label)", $\text{Pred}^{\text{S}}_{\text{update}}$ represents prediction after updating with S, IV=Impact Value, P=Positive-impact, N=Negative-impact.

		MUC	B^3	CEAF_e	CoNLL
Cattan	Gold	83.5	82.4	77.0	81.0
	Predicted	65.9	53.0	44.3	54.4
Ours	Gold	90.5	89.4	86.7	**88.9**
	Predicted	72.8	62.2	53.5	**62.8**

Table 8: ECB+ performance comparison under the setting of extracted mentions.

mentation utilizes the Adam optimizer across all stages.

As shown in Table 9 (using ECB+ as an example), the ablation study on the two key hyperparameters reveals the following patterns:

1) Fixing K=3: Our chosen hidden dimension of 256 is optimal. Reducing it to 128 causes a significant performance drop (-1.4 CoNLL-F1), while increasing it to 512 yields no improvement, indicating 256 is a sufficient and efficient choice.

2) Fixing Hidden Dim=256: Our choice of K=3 is also optimal. Using fewer layers (K=2) leads to insufficient message passing (-1.6 CoNLL-F1), while more layers (K=4) result in over-smoothing and a performance drop.

	Hidden Dim	K	CoNLL
ECB+	128|256|512	3	87.5|**88.9**|88.7
	256	2|3|4	87.3|**88.9**|88.1
GVC	128|256|512	3	86.8|**87.7**|87.6
	256	2|3|4	86.9|**87.7**|87.2
WEC	128|256|512	3	65.6|**66.7**|66.4
	256	2|3|4	65.8|**66.7**|66.4
ECB+META1	128|256|512	3	71.4|**72.2**|71.8
	256	3|4|5	71.1|**72.2**|72.0
ECB+METAm	128|256|512	3	55.8|**56.6**|56.3
	256	3|4|5	55.5|**56.6**|56.5

Table 9: Performance using different hyperparameters.

TR Backbone	CR Backbone	CoNLL-F1
BERT-Large	LongFormer	88.9
DeB3-Large	LongFormer	88.8
MoB-Large	LongFormer	88.9
BERT-Large	DeB3-Large	87.5
MoB-Large	DeB3-Large	87.9
BERT-Large	MoB3-Large	89.1
DeB3-Large	MoB-Large	89.1
MoB-Large	MoB-Large	**89.4**

Table 10: ECB+ results on different combination of backbones. TR=Temporal Relation, CR=Coreference Resolution, DeB3=DeBERTA-v3, MoB=ModernBERT.

H Details of Existing Baselines

CD-DCE proposed by Chen et al. (2025a) enhanced semantics coherence between event contexts to improve CDECR; Metaphoric paraphrasing(Ahmed et al., 2024b) (MP) is the research that proposed ECB+META dataset. Nath et al. (2024) implemented knowledge distillation (KD) methods for event coreference scoring. Min et al. (2024) combined the strengths of LLMs and SLMs (LLM-Min), leading to significant performance improve-

ments. Gao et al. (2024) enhanced CDECR by discourse structure and semantic information (DSSI). Chen et al. (2023) constructed cross-document discourse rhetoric structure (CD-DRS) to capture the global interaction between event mentions.

Additionally, we implement a temporal constraint baseline HT[5] that relies on the HeidelTime temporal normalizer for comparison. HT achieves CoNLL-F1: 77.2 on ECB+, which is significantly lower than our CohTP's 88.9 (-11.7 points). The performance gap underscores that rule-based temporal normalization, while useful, is insufficient for robust CDECR due to limited coverage and inability to capture implicit temporal relationships. Our data-driven approach effectively overcomes these limitations.

I Efficiency Comparison: CohTP vs. CD-DCE

We provide a detailed breakdown using a sample of document cluster from the ECB+ dataset containing approximately 400 event mention pairs and all experiments were conducted on the same NVIDIA RTX 4090 GPU to ensure a fair comparison.

CohTP	$\text{Cost}_{\text{CohTP}}$(s)	$\text{Cost}_{\text{CD-DCE}}$(s)	CD-DCE
TRI	41.2	98.5	DP
TGR	59.4	123.6	TIC
RCC+TP	33.5	252.0	Coh(·)
CTCR	183.6		
TCA	59.1	153.8	MFR
CP	64.8	198.6	CP
Total cost	441.6	826.5	Total cost

Table 11: Cost details of CohTP and CD-DCE, where TRI=Temporal Relation Initialization, TGR=Temporal Graph Refinement, RCC=Residual Conflict Correction, TP=Temporal Partitioning, CTCR=Cross-Time Coref. Recovery, TCA=Temporal Context Augmentation, CP=Coreference Prediction, DP=Document Proprocessing, TIC=Training Instance Construction, MFR=Mention Feature Representation.

From Table 11, we can infer that the CTCR step, which invokes GPT-4 for only 32% of samples, takes 183.6s. If applied to all samples, this step would scale to 573.8s, increasing CohTP's total time to 827.2s, which would exceed CD-DCE's total 822.0s. This demonstrates that our selective intervention strategy is crucial for maintaining the efficiency edge.

[5]1) Using HeidelTime for explicit temporal expression extraction; 2) Applying TF-IDF similarity to find the most relevant document within the same topic for fallback temporal anchors; 3) Ensuring 100% temporal coverage for all event mentions.

J Experimental Results on All Metrics

J.1 Baseline Comparison

Table 12 shows the performance comparison of complete metrics with our CohTP with existing baselines. CohTP outperforms these baselines on all metrics.

ECB+	MUC	B^3	CEAF_e	CoNLL
CD-DCE	90.3	89.0	86.3	88.5
KD	87.9	86.8	84.5	86.4
LLM-Min	88.2	87.8	84.1	86.7
DSSI	86.6	85.4	81.3	84.4
CD-DRS	88.3	87.3	83.6	86.4
HT	81.1	77.3	73.3	77.2
CohTP	**90.5**	**89.4**	**86.7**	**88.9**
GVC	MUC	B^3	CEAF_e	CoNLL
CD-DCE	93.2	86.5	82.3	87.3
KD	92.9	84.3	71.7	83.0
LLM-Min	92.8	87.2	82.1	87.4
HT	79.7	72.4	67.7	73.4
CohTP	**93.5**	**87.3**	**82.4**	**87.7**
WEC	MUC	B^3	CEAF_e	CoNLL
CD-DCE	82.5	67.1	48.2	65.9
DSSI	81.8	65.8	47.3	65.0
HT	66.5	51.0	28.0	48.5
CohTP	**83.1**	**68.0**	**49.1**	**66.7**
ECB+META1	MUC	B^3	CEAF_e	CoNLL
CD-DCE	74.1	73.2	68.7	72.0
MP	-	-	-	71.4
CD-DRS	70.1	72.0	65.5	69.2
HT	58.5	55.0	55.1	56.2
CohTP	**74.2**	**73.5**	**68.9**	**72.2**
ECB+METAm	MUC	B^3	CEAF_e	CoNLL
CD-DCE	58.8	57.3	52.4	56.2
MP	-	-	-	55.6
CD-DRS	54.3	52.8	48.3	51.8
HT	53.2	49.9	47.2	50.1
CohTP	**59.1**	**57.9**	**52.7**	**56.6**

Table 12: Complete metrics of our model and existing baselines.

J.2 LLM Intervention Results

Table 13 shows the complete metrics of different LLM intervention strategies. Our strategy performs best among other strategies across all metrics.

ECB+	MUC	B^3	$CEAF_e$	CoNLL
All Samples	87.1	85.6	82.9	85.2
No Intervention	84.9	78.5	74.6	79.3
CohTP	**90.5**	**89.4**	**86.7**	**88.9**
GVC	MUC	B^3	$CEAF_e$	CoNLL
All Samples	84.5	80.2	77.7	80.8
No Intervention	77.8	69.6	64.4	70.6
CohTP	**93.5**	**87.3**	**82.4**	**87.7**
WEC	MUC	B^3	$CEAF_e$	CoNLL
All Samples	79.5	63.0	45.0	62.5
No Intervention	77.0	62.5	44.7	61.4
CohTP	**83.1**	**68.0**	**49.1**	**66.7**
ECB+META1	MUC	B^3	$CEAF_e$	CoNLL
All Samples	67.5	65.8	62.0	65.1
No Intervention	61.0	58.5	55.1	58.2
CohTP	**74.2**	**73.5**	**68.9**	**72.2**
ECB+METAm	MUC	B^3	$CEAF_e$	CoNLL
All Samples	52.5	50.0	46.3	49.6
No Intervention	47.0	44.5	40.8	44.1
CohTP	**59.1**	**57.9**	**52.7**	**56.6**

Table 13: Results of LLM Intervention Analyses

ECB+	MUC	B^3	$CEAF_e$	CoNLL
LLaMa-Only	77.9	70.8	65.5	71.4
LLaMa+Temp	82.7	75.2	72.6	76.8
CohTP	**90.5**	**89.4**	**86.7**	**88.9**
GVC	MUC	B^3	$CEAF_e$	CoNLL
LLaMa-Only	71.8	64.9	60.5	65.7
LLaMa+Temp	78.1	71.1	64.3	71.2
CohTP	**93.5**	**87.3**	**82.4**	**87.7**
WEC	MUC	B^3	$CEAF_e$	CoNLL
LLaMa-Only	71.5	55.0	39.7	55.4
LLaMa+Temp	73.0	56.5	44.5	58.0
CohTP	**83.1**	**68.0**	**49.1**	**66.7**
ECB+META1	MUC	B^3	$CEAF_e$	CoNLL
LLaMa-Only	66.5	61.8	58.3	62.2
LLaMa+Temp	69.5	65.8	62.7	66.0
CohTP	**74.2**	**73.5**	**68.9**	**72.2**
ECB+METAm	MUC	B^3	$CEAF_e$	CoNLL
LLaMa-Only	51.5	47.0	44.0	47.5
LLaMa+Temp	53.5	49.0	46.6	49.7
CohTP	**59.1**	**57.9**	**52.7**	**56.6**

Table 14: Results of LLM Substitutability Analyses

J.3 LLM Substitutability Results

Table 14 shows the complete metrics regarding LLM substitutability under different settings. Comparing our method with pure LLM-based coreference resolution highlights the critical role of temporal integration. When using LLaMa alone (LLaMa-Only), the model achieves only 71.4 CoNLL-F1 on ECB+, due to the absence of temporal guidance. However, when enhanced with temporal graphs (LLaMa + Temp), the performance improves to 76.8 F1. We employ LLaMa-7B following Chen et al. (2025b), where it demonstrated state-of-the-art performance in direct coreference resolution. This improvement underscores the advantage of dedicated temporal modeling over standalone LLM inference[6].

J.4 Temporal Impact Results

Table 15 shows the complete metrics for ablations on the temporal component, which indicates the effectiveness of combination of temporal and coherence information across all metrics.

J.5 Graph Refinement Results

Table 16 shows the complete metrics for ablations on the graph refinement module. The transitivity component demonstrates the most significant impact among temporal refinement losses[7]. Removing transitivity constraints (w/o Trans) causes CoNLL-F1 drops of 1.5 points on ECB+ (88.9→87.4) and 2.2 points on GVC (87.7→85.5), representing the largest performance degradation across all ablation settings. This substantial decline occurs because transitivity violations directly propagate errors through event chains - a single incorrect relation can corrupt multiple downstream predictions. The symmetry and semantic constraints show milder effects, with F1 reductions of 0.6-1.6 points across datasets. These results confirm that maintaining transitive closure is fundamental to temporal consistency, as it ensures coherent event ordering across multi-hop connections in the graph.

K Graph Refinement Performance

We report the temporal conflict rates before and after EA-GNN refinement across the five evaluation datasets in Table 17. The EA-GNN con-

[6]The prompts for event coreference with and without temporal information is available in Appendix L.2 and L.3 respectively.

[7]The temporal conflict rates before and after EA-GNN refinement across the four evaluation datasets are available in Appendix K.

ECB+	MUC	B^3	$CEAF_e$	CoNLL
w/o Temporal	88.5	87.2	84.1	86.6
w/o Temporal&Coh	86.0	84.5	81.2	83.9
CohTP	**90.5**	**89.4**	**86.7**	**88.9**
GVC	MUC	B^3	$CEAF_e$	CoNLL
w/o Temporal	91.5	85.5	80.4	85.8
w/o Temporal&Coh	88.5	81.5	71.2	80.4
CohTP	**93.5**	**87.3**	**82.4**	**87.7**
WEC	MUC	B^3	$CEAF_e$	CoNLL
w/o Temporal	81.5	65.8	48.0	65.1
w/o Temporal&Coh	80.2	64.5	46.1	63.6
CohTP	**83.1**	**68.0**	**49.1**	**66.7**
ECB+META1	MUC	B^3	$CEAF_e$	CoNLL
w/o Temporal	72.5	71.8	66.0	70.1
w/o Temporal&Coh	71.0	70.5	63.7	68.4
CohTP	**74.2**	**73.5**	**68.9**	**72.2**
ECB+METAm	MUC	B^3	$CEAF_e$	CoNLL
w/o Temporal	57.5	56.2	49.2	54.3
w/o Temporal&Coh	55.5	54.2	41.8	50.5
CohTP	**59.1**	**57.9**	**52.7**	**56.6**

Table 15: Results of Temporal Impact Analyses

sistently reduced temporal conflicts by 58.8% to 66.9%, with the highest reduction on ECB+ (from 24.5% to 8.1%) and the lowest on METAm (from 33.5% to 13.8%). These results demonstrate the robust effectiveness of our graph refinement approach in improving temporal consistency across diverse datasets.

L LLMs Prompts

L.1 Prompt for Temporal Relation Verification

Verify the temporal relationship between two events.

Event A: "{event_a_context}"

Event B: "{event_b_context}"

Current Prediction: {predicted_relation} (confidence: {confidence_score:.2f})

Supporting evidence from nearby events:

- {adjacent_event_1}

- {adjacent_event_2}

Instructions:

1. Analyze the temporal relationship between Event A and Event B

2. Consider the supporting evidence from adjacent events

3. Output ONLY one of: "BEFORE", "SAMETIME", or "UNCLEAR"

ECB+	MUC	B^3	$CEAF_e$	CoNLL
w/o Trans	89.0	88.0	85.2	87.4
w/o Sym	89.5	88.7	86.1	88.1
w/o Sem	89.7	88.9	86.3	88.3
CohTP	**90.5**	**89.4**	**86.7**	**88.9**
GVC	MUC	B^3	$CEAF_e$	CoNLL
w/o Trans	91.5	85.5	79.5	85.5
w/o Sym	92.2	86.2	80.0	86.1
w/o Sem	92.4	86.3	79.9	86.2
CohTP	**93.5**	**87.3**	**82.4**	**87.7**
WEC	MUC	B^3	$CEAF_e$	CoNLL
w/o Trans	80.5	65.5	43.3	63.1
w/o Sym	81.5	66.5	46.1	64.7
w/o Sem	81.7	66.7	46.3	64.9
CohTP	**83.1**	**68.0**	**49.1**	**66.7**
ECB+META1	MUC	B^3	$CEAF_e$	CoNLL
w/o Trans	72.5	71.8	66.9	70.4
w/o Sym	73.0	72.3	67.7	71.0
w/o Sem	72.9	72.2	67.6	70.9
CohTP	**74.2**	**73.5**	**68.9**	**72.2**
ECB+METAm	MUC	B^3	$CEAF_e$	CoNLL
w/o Trans	57.0	55.8	45.9	52.9
w/o Sym	57.5	56.3	46.1	53.3
w/o Sem	57.8	56.6	46.4	53.6
CohTP	**59.1**	**57.9**	**52.7**	**56.6**

Table 16: Results of Graph Refinement Analyses

4. Do not include any explanations or additional text

Output:

L.2 Prompt for Event Coreference Resolution

Determine if the following two event mentions refer to the same real-world event.

Event Mention 1: "{mention_1_text}" Context 1: "{mention_1_context}"

Event Mention 2: "{mention_2_text}" Context 2: "{mention_2_context}"

Instructions:

- Analyze whether these two mentions describe

Dataset	Conflict Rate	Reduction
ECB+	24.5%|8.1%	66.9%
GVC	31.2%|11.7%	62.5%
WEC	33.6%|12.8%	61.9%
ECB+META1	28.8%|10.9%	62.2%
ECB+METAm	33.5%|13.8%	58.8%

Table 17: Temporal conflict rate on average before and after (beforelafter) EA-GNN refinement.

the same specific occurrence

- Consider semantic similarity, arguments, and contextual clues
- Output ONLY "YES" or "NO"
- Do not provide explanations

Answer:

L.3 Prompt for Temporal-aware Event Coreference Resolution

Determine coreference considering temporal constraints.

Event Mentions:

- Mention 1: "{mention_1_text}" (Time: {time_segment_1})
- Mention 2: "{mention_2_text}" (Time: {time_segment_2})

Temporal Graph Information: {relevant_temporal_relations}

Instructions:

1. Check if events are temporally compatible (same or adjacent time segments)
2. Analyze semantic similarity and coreference clues
3. Respect temporal constraints - events in distant segments cannot be coreferent
4. Output ONLY "COREFERENT" or "NON-COREFERENT"

Answer: