Conformal Event Prediction with Temporal Knowledge Graph

Cheng Hu[1,2], Cong Cao[1,2,*], Fangfang Yuan[1,2], Diandian Guo[1,2], Pin Xu[1,2], Yu Liu[1,2], Yanbing Liu[1,2,*]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{hucheng, caocong}@iie.ac.cn

Abstract

Event prediction plays a critical role in high-stakes applications such as military operations, public safety, and healthcare. Current methods learn temporal knowledge graphs to predict events at future timestamps, and the predictions directly influence decision-making and resource allocation. However, these methods lack rigorous uncertainty quantification, which limits their reliability for decision-making, especially in high-stakes scenarios where the cost of errors is high. In this paper, we propose CFEP, a conformal prediction framework tailored for event prediction to address this challenge. This is achieved through end-to-end optimization that ensures coverage while improving efficiency. Specifically, we first introduce non-conformity score diffusion, which captures both topological and temporal uncertainty in temporal knowledge graphs. Additionally, we propose an efficiency-aware optimization algorithm to reduce the coverage gap and improve computational efficiency. Experimental results on three public datasets demonstrate that our approach consistently guarantees statistical coverage while improving efficiency. The code and datasets are available at https://github.com/hucheng-IIE/CFEP.

1 Introduction

Event prediction determines which event types will occur at a future time by learning from historical events, which are extracted from past news articles (Han and Ning, 2022). Event prediction has broad applications in criminal activities (Jhee et al., 2023), political (Deng et al., 2024a), and economic domains (Liu et al., 2024). Current event prediction methods learn temporal knowledge graphs to generate point predictions of future events (Deng et al., 2019; Ma et al., 2023; Deng et al., 2020; Han and Ning, 2022), as illustrated in Fig. 1a. As event prediction is increasingly deployed in high-stakes domains, understanding the uncertainty associated with its predictions becomes crucial.

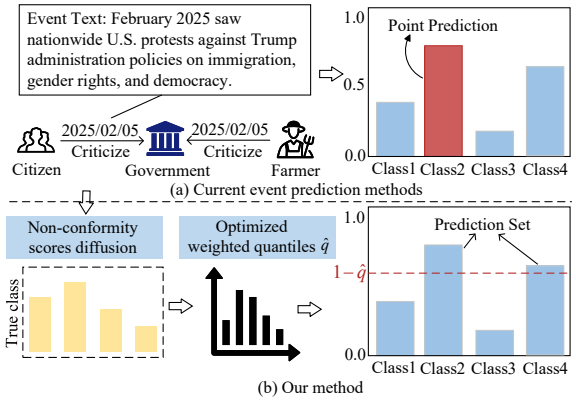

Figure 1: Comparison of our method with current event prediction methods.

To achieve robust uncertainty quantification, researchers have proposed a variety of approaches, including Bayesian-based (Wu et al., 2021), Frequentist-based (Kan et al., 2022), and conformal prediction methods (Vovk et al., 2005a). Among these approaches, conformal prediction stands out due to its distribution-free nature and its ability to provide rigorous statistical guarantees on prediction confidence. The fundamental assumption of conformal prediction is the exchangeability condition [1]. This assumption holds broadly in domains such as computer vision and natural language processing due to the independent nature of data samples. The inherent interdependence of nodes and edges induces structural dependencies that violate the assumption of independent and identically distributed observations, complicating the application of conformal prediction to graph data. Recent work (Huang et al., 2023) addresses this issue by leveraging the permutation equivariant nature of graph

*Corresponding authors

[1]Exchangeability definition: For any $z_1, \ldots, z_{n+1}$ and any permutation π of $\{1, \ldots, n+1\}$, it holds that $\mathbb{P}\big((Z_{\pi(1)}, \ldots, Z_{\pi(n+1)}) = (z_1, \ldots, z_{n+1})\big) = \mathbb{P}\big((Z_1, \ldots, Z_{n+1}) = (z_1, \ldots, z_{n+1})\big)$

Findings of the Association for Computational Linguistics: ACL 2026, pages 5233–5248
July 2-7, 2026

neural network architectures. However, event prediction based on temporal knowledge graphs violates exchangeability. This violation arises because each event in a temporal knowledge graph may follow a distinct distribution, influenced by the temporal dependencies in the event's structure, the attributes of its participants, and the prediction labels. Consequently, the probability of selecting different calibration sets becomes unequal, breaking the exchangeability condition.

As shown in Fig. 1b, we propose CFEP, a novel conformal event prediction framework designed to address this issue. This is achieved by generating prediction sets that are both coverage-guaranteed and highly efficient. We first prove that the exchangeability condition is violated in event prediction, then develop a theory to quantify the coverage gap between exchangeable and non-exchangeable settings. Our analysis indicates that the weighted quantile and the non-conformity score are the primary factors contributing to this discrepancy. Building on these insights, we introduce CFEP, a conformal prediction algorithm for event prediction that calibrates event prediction models by minimizing the deviation from a predefined coverage level. CFEP consists of two modules: non-conformity score diffusion and efficiency-aware optimization. Non-conformity score diffusion leverages both temporal ordering and topological structure to enhance uncertainty quantification in temporal knowledge graphs. Efficiency-aware optimization reduces the coverage gap while improving efficiency. Our main contributions can be summarized as follows:

- We identify the non-exchangeability challenges that arise when applying conformal prediction to event prediction and formally define the conformal prediction problem in this setting.
- We theoretically analyze the impact of weighted quantiles and non-conformity scores on the coverage gap and introduce CFEP to enhance efficiency while maintaining coverage.
- Experimental results on three public datasets demonstrate that our approach consistently guarantees statistical coverage while improving efficiency.

2 Preliminary

In this section, we first provide an overview of event prediction and conformal prediction and then formulate the problem we aim to address. The notations are summarized in Table 5.

Event Prediction. We model historical events to predict co-occurring future events along with their event types (e.g., consult, civil unrest, appeal).

$$\mathbb{P}(\mathbf{y}^{t+\Delta t}|\mathcal{G}^{t-h+1:t}), \tag{1}$$

where $\mathbf{y}^{t+\Delta t} \in \mathbb{R}^{|\mathcal{R}|}$ denotes the probability of co-occurring event types at future timestamp $t+\Delta t$. $\mathcal{G}^{t-h+1:t}$ denotes the temporal knowledge graph within the time window from $t-h+1$ to t.

Definition 1 ***Temporal Knowledge Graph***. *A temporal knowledge graph $\mathcal{G}$ is a directed graph with multiple relations between graph nodes, each relation with a timestamp. For an event occurring at time t, it is depicted as a timestamped edge in the temporal event graph, formalized as a quintuple (s, r, o, t, c), where $s \in \mathcal{E}$ and $o \in \mathcal{E}$ are the head and tail entities, $r \in \mathcal{R}$ is the event type, and $c \in \mathcal{C}$ denotes an event text, with $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{C}$ being finite sets of entities, event types, and event texts, respectively.*

Conformal Prediction. Given a set of data points $(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_n, \mathbf{y}_n), (\mathbf{X}_{n+1}, \mathbf{y}_{n+1})$ and a desired coverage level $1-\alpha \in (0,1)$, consider a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The prediction set for $\mathbf{X}_{n+1}$ is $C(\mathbf{X}_{n+1}) = \{y : s(\mathbf{X}_{n+1}, \mathbf{y}) \leq q\}$, where q is the $\frac{n-1}{n}(1-\alpha)$-th smallest value in the set $\{\, s(\mathbf{X}_i, \mathbf{y}_i) : i = 1, \ldots, n \,\}$. Under this scoring rule, the resulting prediction set satisfies (Vovk et al., 2005b)

$$\mathbb{P}\big(\mathbf{y}_{n+1} \in C(\mathbf{X}_{n+1})\big) \geq 1-\alpha. \tag{2}$$

Conformal prediction relies on the assumption of exchangeability to guarantee that the prediction set achieves a coverage level of at least $1-\alpha$. Recent studies (Huang et al., 2023; Zargarbashi et al., 2023) show that this assumption holds in static graphs. However, event prediction based on temporal knowledge graphs typically violates the exchangeability condition. Here, let $\mathbf{X}^t$ denote the learned representations of each event type at time t, produced by the event prediction model $f(\mathcal{G}^{t-h+1:t})$, and $F(\mathbf{X}^t)$ denote the corresponding predicted labels. $\mathbf{y}$ represent the corresponding ground-truth labels. The detailed proof is provided in Appendix A.1.

Problem Definition. With the aforementioned notations, we formally define the problem of conformal prediction for event prediction. **Given:** (i) The temporal knowledge graph $\mathcal{G}^{t-h+1:t}$ and the ground-truth $\mathbf{y}^t$; (ii) An event prediction model

$f(\mathcal{G}^{t-h+1:t}) = \mathbf{X}^t$ and the transformation function $F(\mathbf{X}^t) = \hat{\mathbf{y}}$; (iii) A predefined miscoverage level $\alpha \in (0, 1)$. **Find:** A prediction set that ensures the ground-truth value falls within the set with a confidence level of at least $1 - \alpha$, while maintaining high efficiency.

3 Theoretical analysis

To address the non-exchangeability issue in event prediction, we introduce an additional compensation term to quantify the coverage gap between exchangeability and non-exchangeability. The coverage gap refers to the discrepancy between exchangeable and non-exchangeable settings given the calibration and test sets, and is defined as follows:

Definition 2 ***Coverage Gap in Event Prediction.*** *Assume $d^{j_{test}} = (\mathbf{X}^{j_t}, \mathbf{y}^{j_t})$ is a randomly selected data point from the test set D_t, and the calibration set D_c contains n_c data points. Together, the data points in D_c and the single test point from D_t form a set of $n_c + 1$ data points. Let C^{j_t} denote the prediction set for the randomly selected test point. The coverage gap is defined as:*

$$\delta_{gap} = (1 - \alpha) - \mathbb{P}\{\mathbf{y}^{j_t} \subseteq C^{j_t}\}. \tag{3}$$

Using the calibration set, test set, and the definition of the coverage gap, we derive an upper bound on this gap (Barber et al., 2023). This bound highlights that the weight quantiles and the non-conformity score calculation methods are the key factors influencing the theoretical coverage in event prediction. It also guides our method to reduce the coverage gap in order to maintain coverage while improving efficiency.

Definition 3 ***Upper Bound for the Coverage Gap.*** *The coverage gap for a test data point $d^{j_t} = (\mathbf{X}^{j_t}, \mathbf{y}^{j_t}) \in D_t$ is bounded by*

$$\delta_{\text{gap}} \leq \frac{\sum_{k=1}^{n_c} \omega_k \, d_{\text{TV}}(\phi, \phi_k)}{1 + \sum_{k=1}^{n_c} \omega_k}, \tag{4}$$

where d_{TV} denotes the total variation distance (Clarkson and Adams, 1933), and ω_i are user-defined weights chosen to make the upper bound small. Let ϕ denote the set containing all non-conformity scores from the calibration set D_c together with the score of the selected test point d^{j_t}:

$$\phi = (s_i^{t_1}, \ldots, s_i^{t_{n_c}}, s_i^{j_t}). \tag{5}$$

For each k, let ϕ_k be the permutation obtained by swapping the test score $s_i^{j_t}$ with the k-th calibration score $s_i^{t_k}$:

$$\phi_k = (s_i^{t_1}, \ldots, s_i^{t_{k-1}}, s_i^{j_t}, s_i^{t_{k+1}}, \ldots, s_i^{t_{n_c}}, s_i^{t_k}). \tag{6}$$

We provide a detailed proof of Eq.4 in Appendix A.2. Eq.4 provides a unified framework that encompasses both the exchangeable and non-exchangeable settings in conformal prediction, making it particularly suitable for real-world event prediction tasks.

4 Methodology

We propose CFEP, an end-to-end optimization procedure that reduces the upper bound of the coverage gap between the calibration and test sets, thereby improving efficiency while maintaining guaranteed coverage. Fig. 2 illustrates our proposed CFEP framework, which consists of two modules: non-conformity score diffusion and efficiency-aware optimization. The non-conformity score diffusion is designed to produce stable non-conformity scores by integrating both the temporal dynamics and the topological structure of the temporal knowledge graph. After obtaining the diffused non-conformity scores, the efficiency-aware optimization assigns differentiated weights to them in order to improve the efficiency of the prediction sets.

4.1 Non-conformity score diffusion

The associations among events are captured by both the temporal dynamics and the topological structure of the temporal knowledge graph. By incorporating these temporal and topological cues into the computation of non-conformity scores, we encourage similar events to exhibit similar non-conformity scores, thereby enhancing the generalization capability of the scores. This section first explains how to select neighboring events that capture both topological and temporal characteristics, and then describes how the non-conformity scores are diffused through these neighboring events.

Neighbor Events. Neighbor events of an event include both temporal neighbors and topological neighbors. Formally, given an event $e_i^{t_1}$ occurring at timestamp t_1 with event type i, its temporal neighbors are defined as follows:

$$N_{i,t_1}^{tpr} = \left\{ e_j^{t_j} \,\middle|\, f(e_i^{t_1}, e_j^{t_j}) = 1,\ |t_1 - t_j| \leq t_{\text{st}} \right\}, \tag{7}$$

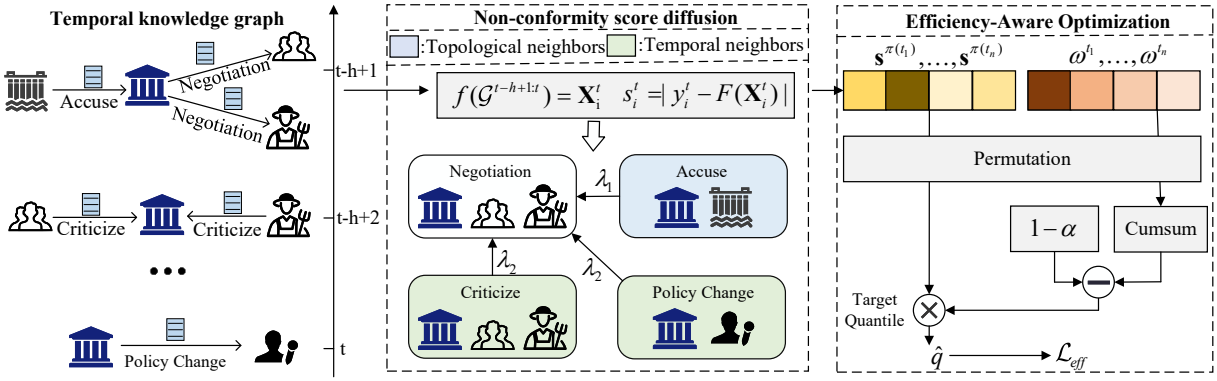

Figure 2: Overview of CFEP. The non-conformity score diffusion incorporates both the temporal information and the topological structure of the temporal knowledge graph to obtain scores with strong generalization. The efficiency-aware optimization is to maintain high efficiency while ensuring the required coverage.

where the function $f(\cdot,\cdot)$ indicates whether two events share a common entity. t_{st} is the predefined temporal range threshold. Topological neighbors N_{i,t_1}^{tpo} refer to events that occur at the same timestamp and share common nodes, which corresponds to $t_{\text{st}} = 0$ in Eq.7. Temporal neighbors and topological neighbors, respectively, represent the temporal information and structural information of events in a temporal knowledge graph. By incorporating these two types of information, robust nonconformity scores can be obtained (Zargarbashi et al., 2023).

Topological and temporal non-conformity scores. The influence of each neighboring event varies; for example, the announcement of a policy may trigger a sequence of subsequent events. To obtain non-conformity scores with strong generalization, we introduce the topological and temporal non-conformity scores defined as follows:

$$s_i^t = \lambda_0 s_i^t + \frac{\lambda_1}{|N_i^{tpo}|} \sum_{j \in N_i^{tpo}} s_j^t + \frac{\lambda_2}{|N_i^{tpr}|} \sum_{k \in N_i^{tpr}} s_k^t, \tag{8}$$

where λ_1 and λ_2 denote how the non-conformity score is influenced by temporal neighbors and structural neighbors, respectively. $\lambda_0 = (1 - \lambda_1 - \lambda_2)$. s_i^t denotes the non-conformity score of the i-th event type at timestamp t. We initialize the diffusion process using the non-conformity score $s_i^t = |y_i^t - F(\mathbf{X}_i^t)|$, where F is the transformation function. Note that each event type i at time t consists of the events occurring within a historical time window, denoted as $E_i^t = \{e_i^{t-h+1}, e_i^{t-h+2}, \ldots, e_i^t\}$. The temporal and topological neighbors of event type i as $N_i^{\text{tpr}} = \bigcup_{\tau=t-h+1}^{t} N_{i,\tau}^{\text{tpr}}$ and $N_i^{\text{tpo}} = \bigcup_{\tau=t-h+1}^{t} N_{i,\tau}^{\text{tpo}}$.

4.2 Efficiency-aware optimization

The nonconformity scores are used to construct the prediction set, and different scores lead to different levels of efficiency. If the prediction set is too inefficient, its applicability in real-world scenarios becomes limited. To reduce the coverage gap between the calibration and test sets, we optimize the parameters ω_i to obtain a highly efficient prediction set while still ensuring the required coverage. To obtain a more precise quantile, we introduce a soft selection mechanism to determine the target quantile, defined as follows:

$$\begin{aligned} \Gamma &= \left| cumsum(\Omega) - (1-\alpha) \right| \\ B &= \left\{ \beta_i \,\middle|\, \beta_i = \frac{e^{-\omega_i/T}}{\sum_{j=1}^{n} e^{-\omega_j/T}} \right\} \\ \hat{q} &= S_\pi B, \end{aligned} \tag{9}$$

where T is a hyperparameter controlling the softness of the prediction set assignment, $\mathbf{S}^\pi = \{\mathbf{s}^{\pi(t_1)}, \ldots, \mathbf{s}^{\pi(t_n)}\}$ denotes the sorted non-conformity scores, and $\Omega = \{\omega^{t_1}, \ldots, \omega^{t_n}\}$ represents the set of weight parameters. The computed quantile serves as the threshold for constructing the prediction set. To improve the efficiency of the prediction set while maintaining the required coverage, we design a loss function to optimize the weighting parameters:

$$\mathcal{L}_{eff} = \sum^{T} \sum^{|\mathcal{R}|} \sigma\left(\frac{s\left(\mathbf{X}_i^t, \mathbf{y}^t\right) - \hat{q}}{\tau} \right), \tag{10}$$

where $s\left(\mathbf{X}_i^t, \mathbf{y}^t\right)$ denotes the nonconformity score for event type i, σ is the sigmoid function, $\hat{q}$ is the differentiable quantile obtained from Eq. 9 and τ is a hyperparameter that controls the sharpness of the assignment.

5 Experiments

In this section, we analyze several key aspects to demonstrate the effectiveness of CFEP: (i) we evaluate CFEP on three benchmark real-world datasets, where CFEP achieves the best overall performance (Section 5.5); (ii) we conduct ablation studies to examine the necessity of each module in CFEP (Section 5.6); (iii) we perform parameter analysis on the nonconformity score diffusion module to justify our parameter choices (Section 5.7); (iv) we carry out robustness analysis to investigate whether CFEP can produce robust nonconformity scores under noisy conditions (Section 5.8); (v) we present a case study to illustrate the superiority of CFEP over existing event prediction models (Section 5.9).

5.1 Dataset

To evaluate the effectiveness of our approach in real-world scenarios, we conduct experiments on the Global Database of Events, Language, and Tone (GDELT) (Leetaru and Schrodt, 2013), a public dataset that offers comprehensive coverage of worldwide information. GDELT extracts events from daily news articles and categorizes them into 20 main categories and multiple subcategories based on the Conflict and Mediation Event Observations (CAMEO) guidelines (Boschee et al., 2015). The dataset records the entities, event type, timestamp, geolocation, and source news text for each event. We select three locations with a high frequency of conflict-related events from the GDELT dataset: Egypt, Iran, and Israel. The time period spans from February 2015 to March 2022, and the temporal granularity is one day. The specific dataset statistics are presented in Table 1. We split the data into train, validation, calibration train, calibration validation, and test datasets in a 5:1:1:1:2 ratio.

Table 1: Dataset Statistics.

Dataset	#urls	$\lvert\mathcal{R}\rvert$	#Events	#Entities	#Days
Egypt	96,081	225	96,081	2,594	2,584
Iran	223,616	236	223,616	2,988	2,584
Israel	345,611	236	345,611	3,456	2,584

5.2 Evaluation Metrics

To rigorously evaluate the performance of CFEP, we adopt two widely used metrics: coverage and efficiency. Coverage measures the reliability of uncertainty estimation by quantifying the proportion of instances in which the prediction set contains the true labels. Efficiency evaluates the precision of the prediction set. The metrics are defined as follows:

$$\text{efficiency} := \frac{1}{|D_t|} \sum_{i \in D_t} \frac{|C_i|}{|Y_i|}, \tag{11}$$

$$\text{coverage} := \frac{1}{|D_t|} \sum_{i \in D_t} \mathbb{I}(y_i \in C_i). \tag{12}$$

where C_i is the prediction set and y_i is the corresponding ground-truth label. We further normalize efficiency by dividing it by the size of the prediction set $|Y_i|$. Larger prediction sets can increase coverage, but they also reduce predictive precision. Striking an appropriate balance between these two metrics is essential for reliable uncertainty quantification. We set the target coverage to 0.95, i.e., $\alpha = 0.05$.

5.3 Baselines

We compare CFEP against several baseline methods. **(1) non-conformity score based methods.** They construct prediction sets by computing different forms of non-conformity scores. We select representative methods: TPS (Sadinle et al., 2019), APS (Romano et al., 2020), RAPS (Angelopoulos et al., 2020) and DAPS (Zargarbashi et al., 2023). **(2) GNN based methods.** We select CF-GNN (Huang et al., 2023), which uses a GNN to calibrate the non-conformity scores. **(3) non-exchangeable based methods.** They address the non-exchangeability problem that arises in graph-structured data. Representative approaches include: NEX (Barber et al., 2023) and NAPS (Clarkson, 2023). **(4) stochastic block model based methods.** UGNN (Davis et al., 2024) samples and aggregates predictions from randomly partitioned graph structures to model uncertainty. To evaluate the effectiveness of these conformal prediction methods, we select four widely used event prediction models: Glean (Deng et al., 2020), MTG (Han and Ning, 2022), TGN (Rossi et al., 2020), and tCompGCN (Vashishth et al., 2020).

5.4 Main Results

The experimental results are summarized in Table 2. CFEP consistently achieves the predefined coverage while maintaining high efficiency across all backbone event prediction models and three public datasets. This demonstrates that the nonconformity score diffusion module effectively generates robust

Table 2: Performance of comparison methods (%). The best results are highlighted in bold, and the second-best results are underlined. ✓ indicates that the method reaches the target coverage (95%) while ✗ indicates the opposite.

Model	Egypt							
	Glean		MTG		TGN		tCompGCN	
	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓
TPS	0.87 ± 0.02✗	4.57 ± 0.11	0.79 ± 0.00✗	5.69 ± 0.01	0.81 ± 0.01✗	4.65 ± 0.03	0.99 ± 0.00✓	7.30 ± 0.02
APS	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	7.33 ± 0.00
RAPS	1.00 ± 0.00 ✓	8.37 ± 0.00	0.68 ± 0.03 ✗	5.12 ± 0.01	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	8.37 ± 0.00
DAPS	0.99 ± 0.01 ✓	4.28 ± 0.22	0.84 ± 0.01 ✗	5.40 ± 0.02	1.00 ± 0.00 ✓	8.37 ± 0.00	0.92 ± 0.00 ✗	5.60 ± 0.00
CF-GNN	0.85 ± 0.06 ✗	4.35 ± 0.01	0.76 ± 0.01 ✗	4.99 ± 0.03	0.66 ± 0.01 ✗	4.92 ± 0.05	0.71 ± 0.00 ✗	5.19 ± 0.05
NEX	0.99 ± 0.01 ✓	4.65 ± 0.04	0.99 ± 0.00 ✓	4.71 ± 0.01	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	7.30 ± 0.02
NAPS	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	8.37 ± 0.00	1.00 ± 0.00 ✓	7.30 ± 0.05
UGNN	0.90 ± 0.00 ✗	5.17 ± 0.16	0.99 ± 0.00 ✓	5.48 ± 0.09	0.97 ± 0.00 ✓	4.26 ± 0.00	0.98 ± 0.00 ✓	3.68 ± 0.01
CFEP	0.97 ± 0.01 ✓	**3.30 ± 0.03**	0.98 ± 0.00 ✓	**3.31 ± 0.00**	0.97 ± 0.01 ✓	**3.13 ± 0.06**	0.96 ± 0.01 ✓	**2.97 ± 0.06**

Model	Iran							
	Glean		MTG		TGN		tCompGCN	
	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓
TPS	0.84 ± 0.01 ✗	5.01 ± 0.01	0.89 ± 0.00 ✗	3.69 ± 0.00	1.00 ± 0.00 ✓	5.37 ± 0.00	0.94 ± 0.00 ✗	3.88 ± 0.01
APS	0.92 ± 0.00 ✗	4.56 ± 0.00	1.00 ± 0.01 ✓	3.06 ± 0.02	1.00 ± 0.00 ✓	5.37 ± 0.00	1.00 ± 0.00 ✓	3.51 ± 0.00
RAPS	1.00 ± 0.00 ✓	5.37 ± 0.00	1.00 ± 0.00 ✓	5.37 ± 0.00	1.00 ± 0.00 ✓	5.37 ± 0.00	1.00 ± 0.00 ✓	5.37 ± 0.00
DAPS	1.00 ± 0.00 ✓	3.80 ± 0.00	1.00 ± 0.00 ✓	3.78 ± 0.01	1.00 ± 0.00 ✓	5.37 ± 0.00	1.00 ± 0.00 ✓	4.75 ± 0.00
CF-GNN	0.79 ± 0.00 ✗	4.16 ± 0.00	0.80 ± 0.00 ✗	4.20 ± 0.00	0.59 ± 0.01 ✗	4.80 ± 0.04	0.79 ± 0.00 ✗	3.16 ± 0.00
NEX	1.00 ± 0.00 ✓	4.02 ± 0.01	1.00 ± 0.00 ✓	4.07 ± 0.04	1.00 ± 0.00 ✓	5.37 ± 0.04	1.00 ± 0.00 ✓	4.00 ± 0.01
NAPS	1.00 ± 0.00 ✓	4.01 ± 0.01	1.00 ± 0.00 ✓	3.07 ± 0.01	1.00 ± 0.00 ✓	4.05 ± 0.01	1.00 ± 0.00 ✓	3.01 ± 0.01
UGNN	1.00 ± 0.00 ✓	5.37 ± 0.01	1.00 ± 0.00 ✓	3.37 ± 0.01	1.00 ± 0.00 ✓	4.23 ± 0.01	0.91 ± 0.00 ✗	5.37 ± 0.00
CFEP	0.98 ± 0.00 ✓	**2.39 ± 0.00**	0.98 ± 0.00 ✓	**2.41 ± 0.00**	0.97 ± 0.00 ✓	**2.48 ± 0.00**	0.98 ± 0.00 ✓	**2.40 ± 0.00**

Model	Israel							
	Glean		MTG		TGN		tCompGCN	
	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓
TPS	0.81 ± 0.00 ✗	4.11 ± 0.01	0.85 ± 0.00 ✗	4.19 ± 0.00	0.84 ± 0.00 ✗	3.65 ± 0.20	0.82 ± 0.00 ✗	3.19 ± 0.01
APS	1.00 ± 0.00 ✓	2.19 ± 0.01	1.00 ± 0.00 ✓	2.19 ± 0.01	1.00 ± 0.00 ✓	3.85 ± 0.01	1.00 ± 0.00 ✓	2.19 ± 0.01
RAPS	1.00 ± 0.00 ✓	3.85 ± 0.00	1.00 ± 0.00 ✓	3.85 ± 0.00	1.00 ± 0.00 ✓	3.85 ± 0.00	1.00 ± 0.00 ✓	3.85 ± 0.00
DAPS	0.92 ± 0.00 ✗	3.14 ± 0.00	1.00 ± 0.00 ✓	2.14 ± 0.00	1.00 ± 0.00 ✓	2.31 ± 0.00	1.00 ± 0.00 ✓	2.14 ± 0.00
CF-GNN	0.84 ± 0.00 ✗	3.17 ± 0.00	0.84 ± 0.00 ✗	3.16 ± 0.00	0.59 ± 0.01 ✗	3.65 ± 0.01	0.84 ± 0.01 ✗	3.17 ± 0.01
NEX	1.00 ± 0.00 ✓	2.14 ± 0.00	1.00 ± 0.00 ✓	2.19 ± 0.01	1.00 ± 0.00 ✓	3.78 ± 0.01	1.00 ± 0.00 ✓	2.19 ± 0.01
NAPS	1.00 ± 0.00 ✓	2.19 ± 0.01	1.00 ± 0.00 ✓	2.19 ± 0.01	1.00 ± 0.00 ✓	3.85 ± 0.01	1.00 ± 0.00 ✓	2.19 ± 0.01
UGNN	1.00 ± 0.00 ✓	3.85 ± 0.01	1.00 ± 0.00 ✓	2.30 ± 0.01	1.00 ± 0.00 ✓	3.85 ± 0.01	1.00 ± 0.00 ✓	3.85 ± 0.01
CFEP	0.97 ± 0.00 ✓	1.75 ± 0.01	0.98 ± 0.00 ✓	1.80 ± 0.00	0.98 ± 0.00 ✓	1.86 ± 0.01	0.98 ± 0.00 ✓	1.80 ± 0.00

nonconformity scores, and the efficiency-aware optimization module reduces the upper bound of the coverage gap by adjusting weighted quantiles, resulting in more compact prediction sets. Notably, nonconformity score-based, GNN-based, and stochastic block model-based methods fail to meet the target coverage across all datasets, reflecting the strong temporal dynamics in event data. Models that perform well under the exchangeability assumption do not generalize effectively to event prediction tasks. While NEX and NAPS attempt to address nonconformity, CFEP achieves significantly higher efficiency. For instance, on the Egypt dataset, CFEP improves efficiency over NEX by 41%, 42%, 167%, and 145% across the four backbone models, highlighting the effectiveness of the efficiency-aware optimization.

5.5 Ablation Study

To verify the effectiveness of the components in CFEP, we conduct an ablation study: (1) Removing temporal neighbors and topological neighbors, denoted as w/o tpr and w/o tpo, respectively. (2) Removing the efficiency-aware loss $\mathcal{L}_{\text{eff}}$, denoted as w/o $\mathcal{L}_{\text{eff}}$. We select the Egypt dataset and use Glean and MTG as the backbone event prediction models. The results are reported in Table 3. In the non-conformity score diffusion module, removing either temporal neighbors or topological neighbors leads to a clear performance degradation, indicating that both temporal and topological uncertainty propagation in temporal knowledge graphs are essential for learning robust nonconformity scores.

Table 3: Ablation Study (%).

Model	Egypt			
	Glean		MTG	
	Coverage ↑	Efficiency ↓	Coverage ↑	Efficiency ↓
w/o tpo	0.96 ± 0.01	3.68 ± 0.00	0.96 ± 0.00	3.56 ± 0.00
w/o tpr	0.95 ± 0.00	3.54 ± 0.00	0.97 ± 0.00	3.49 ± 0.00
w/o $\mathcal{L}_{\text{eff}}$	0.96 ± 0.01	3.70 ± 0.00	0.97 ± 0.00	3.75 ± 0.00
CFEP	**0.97 ± 0.01**	**3.30 ± 0.03**	**0.98 ± 0.00**	**3.31 ± 0.00**

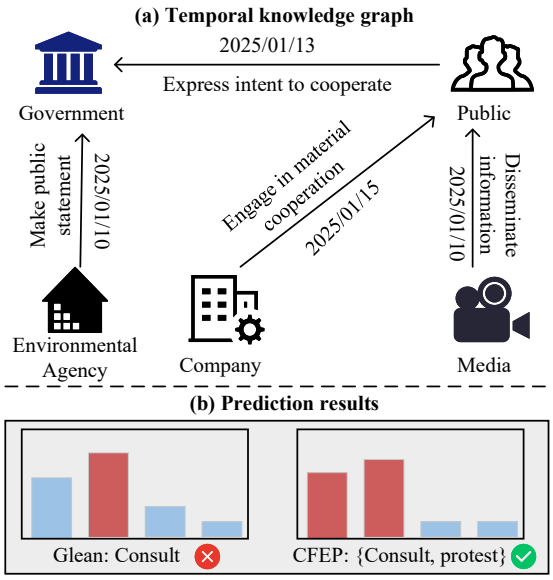

Figure 3: A real sample selected from the test set.

Notably, removing $\mathcal{L}_{\text{eff}}$ leads to a degradation in efficiency, demonstrating that the efficiency-aware optimization module effectively reduces the size of the prediction sets.

Table 4: Parameter Analysis for Non-conformity Score Diffusion.

λ_1	λ_2	Glean		MTG	
		Coverage	Efficiency	Coverage	Efficiency
0.00	0.00	0.95 ± 0.01	1.92 ± 0.01	0.94 ± 0.00	1.87 ± 0.00
0.01	0.01	**0.97 ± 0.01**	**1.75 ± 0.01**	**0.98 ± 0.00**	**1.80 ± 0.00**
0.02	0.01	0.96 ± 0.00	1.85 ± 0.00	0.96 ± 0.00	1.95 ± 0.00
0.03	0.01	0.94 ± 0.01	1.58 ± 0.01	0.94 ± 0.00	1.97 ± 0.00
0.04	0.01	0.94 ± 0.00	1.52 ± 0.00	0.94 ± 0.00	1.82 ± 0.00
0.02	0.02	0.93 ± 0.01	1.46 ± 0.01	0.94 ± 0.00	2.13 ± 0.00
0.03	0.02	0.93 ± 0.01	1.50 ± 0.00	0.93 ± 0.00	1.92 ± 0.00
0.05	0.01	0.93 ± 0.01	1.64 ± 0.01	0.92 ± 0.01	1.53 ± 0.01
0.01	0.05	0.93 ± 0.01	1.61 ± 0.01	0.91 ± 0.01	1.27 ± 0.01
0.1	0.5	0.92 ± 0.01	1.52 ± 0.01	0.91 ± 0.00	1.23 ± 0.00
0.5	0.1	0.91 ± 0.00	1.26 ± 0.00	0.90 ± 0.00	1.16 ± 0.00
1.0	0.0	0.87 ± 0.00	1.16 ± 0.00	0.90 ± 0.00	1.13 ± 0.00
0.0	1.0	0.91 ± 0.00	1.25 ± 0.00	0.90 ± 0.00	1.10 ± 0.00

5.6 Parameter Sensitivity Analysis

We empirically investigate the impact of λ_1 and λ_2 in Eq. 8 on the performance of CFEP. The parameter analysis is conducted on the Israel dataset using Glean as the backbone event prediction model. We choose this dataset because events evolve rapidly over time, making it a representative benchmark for examining the sensitivity of CFEP to parameter variations. The results are reported in Table 4. Specifically, λ_1 controls the contribution of topological neighbors to the nonconformity score, while λ_2 governs the contribution of temporal neighbors. The experimental results indicate that increasing the parameters generally reduces the size of the prediction set, thereby sacrificing coverage. Setting both λ_1 and λ_2 to small values ($\lambda_1 = \lambda_2 = 0.01$) mitigates the influence of topological and temporal neighbors, achieving the optimal balance between coverage and efficiency.

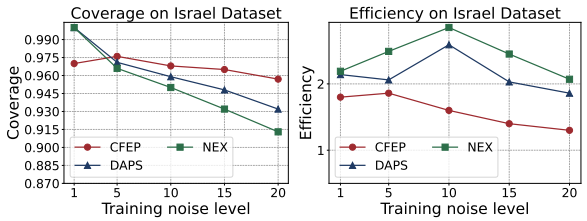

Figure 4: Efficiency and Coverage vs Training noise level in Israel.

5.7 Robustness Analysis

In real-world applications, data inevitably contain noise, and the conformal prediction methods are required to maintain reliable predictions even under extreme conditions. We conduct experiments on the Israel dataset using MTG as the backbone event prediction model and select DAPS and NEX as baseline methods, as they achieve strong performance in the Israel dataset. Poisson noise is injected into the training and calibration data to analyze how coverage and efficiency vary with increasing noise levels. As shown in Fig. 4, CFEP demonstrates substantially stronger robustness than DAPS and NEX. By incorporating both temporal and topological neighbors, CFEP learns robust nonconformity scores that preserve the target coverage while maintaining high efficiency even in noisy scenarios. DAPS exhibits greater robustness than NEX, which can be attributed to its ability to integrate structural information and global relationships. NEX primarily emphasizes explicit reduction of prediction set size, leading to nonconformity scores that are more sensitive to noise.

5.8 Case Study

We select a real test sample to illustrate the difference between CFEP and existing event prediction models. We use Glean, a representative method that performs point predictions by learning temporal knowledge graphs, to represent current event prediction approaches. As shown in Fig. 3 (a), a temporal knowledge graph is constructed from the selected test sample, with the ground truth event at the future timestamp being a protest. Fig. 3 (b)

shows that Glean predicts the event with the highest probability, which is a Consult event. CFEP produces a prediction set {Consult, Protest} that includes the true label. The point predictions produced by Glean do not provide information about prediction confidence, and errors can result in significant losses, making them difficult to use in practical decision-making. CFEP generates prediction sets with guaranteed coverage while maintaining high efficiency, thereby providing substantial support for informed decision-making.

6 Related Work

6.1 Event Prediction

Research on event prediction spans a wide range of domains, including criminal activities (Sun et al., 2023; Jhee et al., 2023; Hu et al., 2024; Qin et al., 2025; LIN et al., 2024), political events (Deng et al., 2024a; Thida, 2026; Gwak et al., 2024; Shahi et al., 2024; von der Maase, 2025), and stock markets (Liu et al., 2024; Koa et al., 2024; Saberironaghi et al., 2025; Chiu et al., 2025). Early approaches to event prediction employ machine learning techniques (Kallus, 2014), which rely on statistical features to predict events. Recently, researchers have used deep learning methods to automatically extract event features (Deng et al., 2024b; Wu et al., 2020). To represent event structures, some researchers model events as graphs and leverage semantic information from event texts to derive rich representations, such as topic-related keywords, document embeddings, causal effects, knowledge graphs (Deng et al., 2021, 2020, 2022b,a), and contextual information (Ma et al., 2023; Han and Ning, 2022). However, none of these methods consider uncertainty quantification for event prediction.

6.2 Conformal Prediction

Vovk (Vovk et al., 2005b, 2017) first offers prediction sets with guaranteed confidence levels based on a predefined coverage rate. Since its inception, numerous studies have focused on both advancing its theoretical foundations (Tibshirani et al., 2019; Xu and Xie, 2021; Campos et al., 2024; Gibbs et al., 2025) and improving practical applications (Fannjiang et al., 2022; Lu et al., 2022; Quach et al., 2023; Mossina et al., 2024; Ernez et al., 2023; Pantelidis et al., 2025; Silva-Rodríguez et al., 2025; Everink et al., 2025). Given its inherent robustness, a key challenge in applying conformal prediction across different domains is enhancing efficiency (Ndiaye, 2022; Straitouri et al., 2023). The design of an effective non-conformity score plays a critical role in improving the efficiency of prediction sets while maintaining exchangeability guarantees. Several notable approaches, including TPS (Sadinle et al., 2019), APS (Romano et al., 2020), and RAPS (Angelopoulos et al., 2020), achieve higher efficiency by employing distinct strategies for computing non-conformity scores. More recently, several studies have extended conformal prediction to graph-structured domains (Huang et al., 2023; Zhao et al., 2024; Zargarbashi and Bojchevski, 2024; Zhang et al., 2025; Song et al., 2024). (Zargarbashi et al., 2023) proposed a diffusion-based non-conformity score that incorporates topological information for graph-based conformal prediction. In contrast to these studies, our work targets the domain of event prediction over temporal knowledge graphs, where the standard exchangeability assumption breaks down.

7 Conclusion

In this paper, we introduce CFEP, a framework that applies conformal prediction to event prediction. Our primary objective is to integrate nonconformity theory into the event prediction framework and demonstrate that the exchangeability assumption is violated in this domain due to the temporal dependencies of events. Through theoretical analysis, we propose an upper bound to effectively address the coverage gap. CFEP incorporates the nonconformity score diffusion module, which simultaneously considers both topological and temporal neighbors to generate robust nonconformity scores, while the efficiency-aware optimization module adjusts the quantile weights to produce prediction sets that achieve high efficiency without compromising coverage. Experiments on three real-world datasets show that CFEP outperforms existing baseline methods in terms of both efficiency and coverage.

Limitations

CFEP inherits the representational limitations of the underlying event prediction model. When the backbone model produces poorly calibrated or noisy outputs, CFEP may yield conservative prediction sets with reduced efficiency. Moreover, CFEP currently assumes a predefined event type space during calibration, which limits its direct applicability to entirely unseen event types. In future work,

we plan to further enhance the representational robustness of CFEP so that it remains effective even when the backbone model provides suboptimal predictions.

Ethics Statement

Regarding the potential risks associated with our work, we provide the following clarifications. This work focuses on developing an uncertainty quantification framework for event prediction and is intended solely for research purposes. It should not be directly used for policy-making or operational decision-making. In the case of processing event text with pre-trained language models, we freeze the model parameters and employ the model in an offline setting to effectively mitigate the risk of data leakage. No AI assistance was applied in any of the experiments. AI was only used to assist in refining the language of the paper to enhance clarity. The novel contributions, motivations, and figures in this work were developed without the aid of AI.

Acknowledgments

This research was supported by the National Key R&D Program of China (No. 2023YFC3303800).

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*, pages 816–845.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. ICEWS Coded Event Data.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

Cheng-Chih Chiu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2025. Pre-finetuning with impact duration awareness for stock movement prediction. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 929–933.

James A Clarkson and C Raymond Adams. 1933. On definitions of bounded variation for functions of two variables. *Transactions of the American Mathematical Society*, pages 824–854.

Jase Clarkson. 2023. Distribution free prediction sets for node classification. In *International conference on machine learning*, pages 6268–6278. PMLR.

Ed Davis, Ian Gallagher, Daniel John Lawson, and Patrick Rubin-Delanchy. 2024. Valid conformal prediction for dynamic gnns. *arXiv preprint arXiv:2405.19230*.

Songgaojun Deng, Maarten de Rijke, and Yue Ning. 2024a. Advances in human event modeling: From graph neural networks to language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6459–6469. ACM.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2019. Learning dynamic context graphs for predicting social events. In *KDD*, pages 1007–1016.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2020. Dynamic knowledge graph based multi-event forecasting. In *KDD*, pages 1585–1595.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2021. Understanding event predictions via contextualized multilevel feature learning. In *CIKM*, pages 342–351.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2022a. Causality enhanced societal event forecasting with heterogeneous graph learning. In *IEEE International Conference on Data Mining, ICDM 2022, Orlando, FL, USA, November 28 - Dec. 1, 2022*, pages 91–100. IEEE.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2022b. Robust event forecasting with spatiotemporal confounder learning. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 294–304. ACM.

Songgaojun Deng, Olivier Sprangers, Ming Li, Sebastian Schelter, and Maarten de Rijke. 2024b. Domain generalization in time series forecasting. *TKDD*, 18(5):1–24.

Fares Ernez, Alexandre Arnold, Audrey Galametz, Catherine Kobus, and Nawal Ould-Amer. 2023. Applying the conformal prediction paradigm for the uncertainty quantification of an end-to-end automatic speech recognition model (wav2vec 2.0). In *Conformal and Probabilistic Prediction with Applications*, pages 16–35. PMLR.

Jasper Marijn Everink, Bernardin Tamo Amougou, and Marcelo Pereyra. 2025. Self-supervised conformal prediction for uncertainty quantification in imaging problems. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 108–118. Springer.

Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. 2022. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, page e2204569119.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. 2025. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf008.

Daehoon Gwak, Junwoo Park, Minho Park, Chaehun Park, Hyunchan Lee, Edward Choi, and Jaegul Choo. 2024. Forecasting future international events: A reliable dataset for text-based event modeling. *arXiv preprint arXiv:2411.14042*.

Xiaoxue Han and Yue Ning. 2022. Text-enhanced multi-granularity temporal graph learning for event prediction. In *IEEE International Conference on Data Mining, ICDM 2022, Orlando, FL, USA, November 28 - Dec. 1, 2022*, pages 171–180. IEEE.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. 2021. Explainable subgraph reasoning for forecasting on temporal knowledge graphs. *International Conference on Learning Representations,International Conference on Learning Representations*.

Kaixi Hu, Lin Li, Qing Xie, Xiaohui Tao, and Guandong Xu. 2024. Crimealarm: Towards intensive intent dynamics in fine-grained crime prediction. In *International Conference on Database Systems for Advanced Applications*, pages 104–120. Springer.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. 2023. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, pages 699–721.

Jong Ho Jhee, Myung Jun Kim, Myeonggeon Park, Jeongheun Yeon, and Hyunjung Shin. 2023. Fast prediction for criminal suspects through neighbor mutual information-based latent network. *IJIS*, pages 1–12.

Nathan Kallus. 2014. Predicting crowd behavior with big public data. In *WWW*, pages 625–630.

Kelvin Kan, François-Xavier Aubet, Tim Januschowski, Youngsuk Park, Konstantinos Benidis, Lars Ruthotto, and Jan Gasthaus. 2022. Multivariate quantile function forecaster. In *International Conference on Artificial Intelligence and Statistics*, pages 10603–10621. PMLR.

Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4304–4315.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone. In *ISA annual convention*, pages 1–49. Citeseer.

Yihang LIN, Kun ZHENG, Shuhao XIA, Li QI, Jie DAI, Xuan CAI, and Qinggang ZHU. 2024. A crime prediction model incorporating regional spatial similarity characteristics and spatio temporal characteristics of events. *Geomatics and Information Science of Wuhan University*.

Mengpu Liu, Mengying Zhu, Xiuyuan Wang, Guofang Ma, Jianwei Yin, and Xiaolin Zheng. 2024. Echo-gl: Earnings calls-driven heterogeneous graph learning for stock movement prediction. In *AAAI*, 12, pages 13972–13980.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. 2022. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12008–12016.

Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. 2023. Context-aware event forecasting via graph disentanglement. In *KDD*, pages 1643–1652.

Luca Mossina, Joseba Dalmau, and Léo Andéol. 2024. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3574–3584.

Eugene Ndiaye. 2022. Stable conformal prediction sets. In *International Conference on Machine Learning*, pages 16462–16479. PMLR.

Ippokratis Pantelidis, Korbinian Randl, and Aron Henriksson. 2025. Efficient text classification with conformal in-context learning. *arXiv preprint arXiv:2512.05732*.

Zhenkai Qin, BaoZhong Wei, and Caifeng Gao. 2025. Innovative lsgtime model for crime spatiotemporal prediction based on mindspore framework. *arXiv preprint arXiv:2503.20136*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, pages 3581–3591.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. In *ICML*, pages 1–16.

Mohammadreza Saberironaghi, Jing Ren, and Alireza Saberironaghi. 2025. Stock market prediction using machine learning and deep learning techniques: A review. *AppliedMath*, 5(3).

Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, pages 223–234.

Gautam Kishore Shahi, Ali Sercan Basyurt, Stefan Stieglitz, and Christoph Neuberger. 2024. Agenda formation and prediction of voting tendencies for european parliament election using textual, social and network features. *Information Systems Frontiers*, pages 1–19.

Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. 2025. Conformal prediction for zero-shot models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19931–19941.

Jianqing Song, Jianguo Huang, Wenyu Jiang, Baoming Zhang, Shuangjie Li, and Chongjun Wang. 2024. Similarity-navigated conformal prediction for graph neural networks. *Advances in Neural Information Processing Systems*, 37:48541–48567.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. 2023. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pages 32633–32653. PMLR.

Yuting Sun, Tong Chen, and Hongzhi Yin. 2023. Spatial-temporal meta-path guided explainable crime prediction. *World Wide Web*, 26(4):2237—-2263.

Myo Thida. 2026. Time-series forecasting for political violence targeting women. *International Journal of Data Science and Analytics*, 21(1):64.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *ICLR*, pages 1–16.

Simon P von der Maase. 2025. Next-generation conflict forecasting: Unleashing predictive patterns through spatiotemporal learning. *arXiv preprint arXiv:2506.14817*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005a. *Algorithmic learning in a random world*. Springer.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005b. Algorithmic learning in a random world. *Springer*, pages 1–9.

Vladimir Vovk, Jieli Shen, Valery Manokhin, and Minge Xie. 2017. Nonparametric predictive distributions based on conformal prediction. In *Conformal and probabilistic prediction and applications*, pages 82–102. PMLR.

Dongxia Wu, Liyao Gao, Matteo Chinazzi, Xinyue Xiong, Alessandro Vespignani, Yi-An Ma, and Rose Yu. 2021. Quantifying uncertainty in deep spatiotemporal forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1841–1851.

Xian Wu, Chao Huang, Chuxu Zhang, and Nitesh V Chawla. 2020. Hierarchically structured transformer networks for fine-grained spatial event forecasting. In *WWW*, pages 2320–2330.

Chen Xu and Yao Xie. 2021. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR.

Soroush H Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. 2023. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning*, pages 12292–12318. PMLR.

Soroush H Zargarbashi and Aleksandar Bojchevski. 2024. Conformal inductive graph neural networks. *arXiv preprint arXiv:2407.09173*.

Zheng Zhang, Jie Bao, Zhixin Zhou, Nicolo Colombo, Lixin Cheng, and Rui Luo. 2025. Residual reweighted conformal prediction for graph neural networks. *arXiv preprint arXiv:2506.07854*.

Tianyi Zhao, Jian Kang, and Lu Cheng. 2024. Conformalized link prediction on graph neural networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4490–4499.

Table 5: Important notations and descriptions.

Notation	Description
s, r, o, c	head entity, event type, tail entity and event text
$\mathcal{E}, \mathcal{R}, \mathcal{C}$	sets of entities, event types and event texts
$F(\cdot)$	the transformation function
$f(\cdot)$	event prediciton model
E	the set of events
e_i^t	the event of type i occurring at timestamp t
$\mathbf{X}_i^t$	the representation of the event type i at timestamp t
s_i^t	the nonconformity score of the event type i at timestamp t

A Theory Analysis

A.1 Analysis of Non-Exchangeability in Event Prediction

Table 5 summarizes the important notations and their corresponding descriptions used in this paper. We theoretically analyze the violation of exchangeability in event prediction. Before presenting the

theoretical framework, we define the calibration and test sets in event prediction.

Definition 4 ***Calibration Set and Test Set.*** *Given the set of events of type i extracted from the temporal knowledge graph, denoted as $E_i^{\leq t} = \{e_i^{t_1}, e_i^{t_2}, \ldots, e_i^t\} \subseteq E$ with timestamps satisfying $t_1 \leq \cdots \leq t$. The nonconformity scores for each event type i are s_i^t. The sets of events in the calibration set and the test set are defined as E_c and E_t, respectively:*

$$D_c = (E_c^{\leq t}, \mathbf{X}_i^t, \mathbf{y}_i^t, s_i^t), E_c^{\leq t} = e_i^{t-h+1:t} \quad (13)$$

$$D_t = (E_t^{\leq t}, \mathbf{X}_j^t, \mathbf{y}_j^t, s_j^t), E_t^{\leq t} = e_j^{t-h+1:t}, \quad (14)$$

where $D_c \cap D_t = \varnothing$ and $E_c \cap E_t = \varnothing$. $\mathbf{X}_i^t$, $\mathbf{y}_i^t$, and s_i^t denote the feature representation, ground-truth label, and nonconformity score of event type i at timestamp t, respectively. $E_c^{\leq t}$ denotes the set of events extracted from the temporal knowledge graph $\mathcal{G}^{t-h+1:t}$.

We begin by establishing that exchangeability holds for event prediction on static graphs, followed by an analysis showing that it is violated in temporal knowledge graphs. In a static graph $G = (V, E)$, given a training dataset D_{t}, a validation dataset D_{v}, a calibration dataset D_{c} and a test dataset D_{t}, if a model mapping data points X to Y and a function mapping $\mathcal{X} \times \mathcal{Y}$ satisfy the assumption in Eq. (15), then the exchangeability condition can be maintained:

$$S\Big(x, y; \{x_i, y_i\}_{i \in D_{\text{t}} \cup D_{\text{v}}}, \{x_j\}_{j \in D_{\text{c}} \cup D_{\text{t}}}\Big) = S\Big(x, y; \{x_i, y_i\}_{i \in D_{\text{t}} \cup D_{\text{v}}}, \{x_{\pi(j)}\}_{j \in D_{\text{c}} \cup D_{\text{t}}}\Big), \quad (15)$$

where π denotes a permutation over the events in the calibration and test sets. S is the non-conformity score function. In general, event prediction models satisfy Eq. (15) because they depend solely on the graph structure and event attributes and do not take the temporal order into account. In a temporal knowledge graph, the graph topology, the labels of event types, and the features of events all evolve. The violation of exchangeability arises for two reasons. First, the nonconformity scores of the calibration and test sets inherently depend on time, causing them to follow different distributions. Second, training an event prediction model requires temporal information, meaning that the ordering of nodes and edges in the temporal graph influences the model's predictions. Both factors result in a violation of Eq. (15).

Proposition 1 ***Non-Exchangeability in Event Prediction.*** *Suppose there exists an event type i such that $(s_i^{t_1}, \ldots, s_i^{t_k}) \sim P_t$ and $(s_i^{t_{k+1}}, \ldots, s_i^{t_n}) \sim P_{t+\Delta t}$. Then, the probability of selecting n_c nonconformity scores to form the calibration set can be written as*

$$P(E_c \mid E_{c_t}) = \prod_{j=1}^{k} P_t(s_i^{t_j}) \prod_{j=k+1}^{n} P_{t+\Delta t}(s_i^{t_j}),$$

and for any permutation π of the indices, it holds that

$$P(E_c \mid E_{c_t}) \neq P(E_{\pi(c)} \mid E_{c_t}).$$

Despite this, existing studies have not examined conformal prediction in the context of event prediction.

A.2 Proof for the Upper Bound on Coverage under Non-Exchangeability

Given a calibration set D_c and a randomly selected test instance from D_t, we compute the non-conformity scores for all samples in the calibration set as well as for the test instance. Let $\phi_c = (s^{t_1}, \ldots, s^{t_{n_c}})$ denote the set of non-conformity scores obtained from the calibration set D_c. Each s^t is computed as the average of non-conformity scores across all event types at time t:

$$s^t = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} s_i^t, \quad (16)$$

where $\mathcal{R}$ represents the set of event types, and s_i^t is the non-conformity score of event type i at time t. When incorporating the j-th test sample from D_t, we define $\phi_j = (s^{t_1}, \ldots, s^{t_{n_c}}, s^{j_t})$ as the augmented set of non-conformity scores. Furthermore, we use $\phi_j^k = (s^{t_1}, \ldots, s^{t_{k-1}}, s^{j_t}, s^{t_{k+1}}, \ldots, s^{t_{n_c}}, s^{t_k})$ to represent a permutation in which the test score s^{t_k} is swapped with the k-th calibration score.

According to conformal prediction theory, when the non-conformity scores fail to satisfy the exchangeability assumption, the prediction event can be characterized as:

$$y^{j_t} \notin C^{j_t} \iff s^{j_t} > Q_{1-\alpha}\left(\sum_{i=1}^{n_c+1} \tilde{\omega}_i \cdot \delta_{\Phi_j}\right), \quad (17)$$

where $\delta(\cdot)$ denotes sorting the non-conformity scores in ascending order. Here, $\tilde{\omega}_i$ are predefined weights satisfying $\sum_{i=1}^{n_c+1} \tilde{\omega}_i = 1$. Under the exchangeability assumption, these weights reduce to uniform values, i.e., $\tilde{\omega}_i = 1/(n_c + 1)$.

Eq. 17 illustrates the effect of violating exchangeability. To preserve coverage guarantees, it is therefore sufficient to show that the quantile computed from any permuted set $\Phi_j^{(k)}$ does not exceed that derived from Φ_j. Formally, we aim to establish

$$Q_{1-\alpha}\left(\sum_{i=1}^{n_c}\tilde{\omega}_i\cdot\delta_{\Phi_j}+\tilde{\omega}_{n_c+1}\delta_{n_c+1}\right) \geq Q_{1-\alpha}\left(\sum_{i=1}^{n_c+1}\tilde{\omega}_i\cdot\delta_{\Phi_j^{(k)}}\right). \tag{18}$$

From this, it is evident that the inequality holds when the test instance has the largest non-conformity score. Thus, it remains to verify the condition when the test score is not the maximum. Consider a specific permutation $\Phi_j^{(k)}$. The corresponding weighted quantile can be expressed as:

$$\sum_{i=1}^{k-1}\tilde{\omega}_i s^{t_i}+\tilde{\omega}_k s^{t_{n_c+1}}+\sum_{i=k+1}^{n_c}\tilde{\omega}_i s^{t_i}+\tilde{\omega}_{n_c+1}s^{t_k}. \tag{19}$$

Meanwhile, the quantile computed from Φ_j can be rewritten as

$$\sum_{i=1}^{k-1}\tilde{\omega}_i s^{t_i}+\tilde{\omega}_k s^{t_k}+\sum_{i=k+1}^{n_c}\tilde{\omega}_i s^{t_i}+\tilde{\omega}_{n_c+1}s^{t_{n_c+1}}. \tag{20}$$

To ensure that the inequality holds for any permutation $\Phi_j^{(k)}$, it suffices to show that the quantile from Φ_j is no smaller than that from $\Phi_j^{(k)}$. Subtracting the two expressions yields $(\tilde{\omega}_{n_c+1}-\tilde{\omega}_k)(s^{t_{n_c+1}}-s^{t_k})$. For this to be non-negative, it is necessary that $\tilde{\omega}_{n_c+1}\geq\tilde{\omega}_k$, given that exchangeability violation implies $s^{t_{n_c+1}}>s^{t_k}$, where s^{t_k} corresponds to the score responsible for the violation. Consequently, $y^{j_t}\notin C^{j_t}\ \Rightarrow\ s^{t_k}\in\Phi_j^{(k)}$. This leads to

$$\mathbb{P}(y^{j_t}\notin C^{j_t})=\mathbb{P}(s^{t_k}\in\Phi_j^{(k)})= \sum_{i=1}^{n_c+1}\tilde{\omega}_i\cdot\mathbb{P}(i\in\Phi_j^{(k)}). \tag{21}$$

By applying the total variation distance, we further obtain

$$\mathbb{P}(y^{j_t}\notin C^{j_t})\leq\sum_{i=1}^{n_c+1}\tilde{\omega}_i\Big(\mathbb{P}(i\in\Phi_j)+d_{\mathrm{TV}}(\Phi_j,\Phi_j^{(k)})\Big), \tag{22}$$

which can be rewritten as

$$\mathbb{E}\left[\sum_{i\in\Phi_j}\tilde{\omega}_i\right]+\sum_{i=1}^{n_c}\tilde{\omega}_i\cdot d_{\mathrm{TV}}(\Phi_j,\Phi_j^{(k)}) \leq\alpha+\sum_{i=1}^{n_c}\tilde{\omega}_i\cdot d_{\mathrm{TV}}(\Phi_j,\Phi_j^{(k)}). \tag{23}$$

B Methodology

B.1 Algorithm of CFEP

The overall procedure of the proposed method is summarized in Algorithm 1. We first train the event prediction model on the training and validation sets. The trained model is then used to obtain event representations as well as the corresponding logits, which serve as the basis for computing the base non-conformity scores. Next, CFEP is trained on the calibration set. During the Non-Conformity Score Diffusion stage, we aggregate both topological neighbors and temporal neighbors for each event category, resulting in temporally and topologically diffused non-conformity scores for each event. After obtaining the diffused non-conformity scores across the entire calibration set, we estimate the quantile threshold $\hat{q}$ through a differentiable procedure proposed in this work and further optimize a weighted quantile in the Efficiency-aware Optimization stage. During inference on the test set, we construct the prediction set by comparing the non-conformity scores with the weighted quantile obtained by CFEP, thereby producing prediction sets that guarantee coverage while achieving high efficiency.

B.2 Time Complexity

The time complexity of the CFEP method can be analyzed by considering the key steps involved in both the training and inference phases.

First, the event prediction model is trained on the temporal knowledge graph. This step typically involves learning a model that captures the temporal and relational dependencies between events. The training time for this model depends on the specific architecture used (e.g., Graph Neural Networks, LSTM, or Transformer models) and is denoted as $O(T_{\text{train}})$, where T_{train} is the time complexity of the prediction model training process. Depending on the model complexity, this can vary, but it is generally proportional to the size of the graph and the number of training epochs.

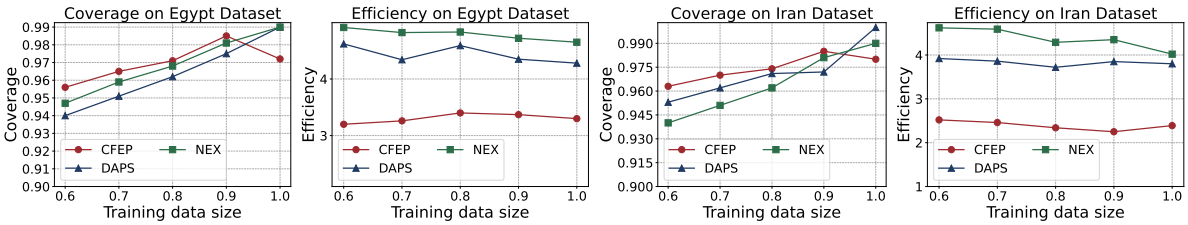

Figure 5: Efficiency and Coverage vs Training Data Size on Egypt and Iran.

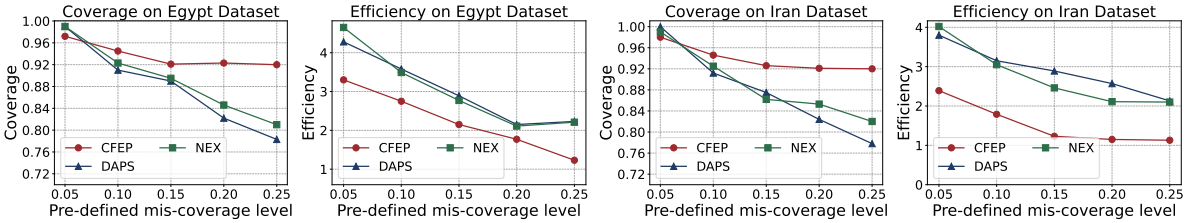

Figure 6: Efficiency and Coverage vs Pre-defined mis-coverage level on Egypt and Iran.

Second, for each event type $i \in \mathcal{R}$, the non-conformity score s_i^t is calculated. This process involves comparing the predicted event labels with the true labels, which can be done in $O(n)$ time, where n is the number of event types. Each non-conformity score captures how "unexpected" or "non-conforming" a given prediction is, and this calculation is done individually for each event type in the dataset.

Once the non-conformity scores are computed, the next step is to diffuse these scores across the temporal and topological neighbors of each event. The diffusion process involves propagating the non-conformity scores through the graph, taking into account both temporal relationships (i.e., events occurring within a certain time window) and topological relationships (i.e., events that are connected through shared entities or relationships). The time complexity of this operation is proportional to the number of edges in the graph, $O(|E|)$, where $|E|$ is the number of edges in the temporal knowledge graph. This step typically involves a message-passing or graph traversal mechanism, which ensures that each event's score is updated based on its neighbors.

Finally, the quantile threshold for conformal prediction is computed by sorting the non-conformity scores. This sorting step ensures that the prediction set can be constructed according to the desired coverage level $1-\alpha$. Sorting n non-conformity scores requires $O(n \log n)$ time, which is the dominant complexity in the quantile computation.

Therefore, combining all the steps, the overall time complexity of CFEP is given by the following expression:

$$T_{\text{CFEP}} = O(T_{\text{train}}) + O(n \cdot k) + O(|E|) + O(n \log n) \tag{24}$$

where T_{train} is the training time of the event prediction model, n is the number of event types, k represents the average number of neighbors per event, $|E|$ is the number of edges in the temporal knowledge graph, and $n \log n$ accounts for the sorting of non-conformity scores.

C Experiments

C.1 Dataset

The majority of existing benchmark datasets are constructed from different curated subsets of GDELT and ICEWS. For instance, the commonly used ICEWS14 and ICEWS18 datasets represent event records extracted from the ICEWS corpus (Han et al., 2021) for the years 2014 and 2018, respectively. Nevertheless, these datasets do not preserve the original news articles from which the events are derived. Instead, they only retain structured event attributes, including the head entity, tail entity, event type, and timestamp. The omission of raw news text substantially limits semantic analysis, as textual context plays a critical role in capturing event semantics and interpreting the underlying meanings of events. Building upon the SeCoGD (Ma et al., 2023), we further perform additional data cleaning to remove noisy records from the dataset.

Algorithm 1 CFEP: Conformal Event Prediction with Temporal Knowledge Graph

Require: Temporal knowledge graph $\mathcal{G}^{t-h+1:t}$; event prediction model $f(\cdot)$; transformation function $F(\cdot)$; train set $\mathcal{D}_{\text{train}}$, valid set $\mathcal{D}_{\text{valid}}$, calibration set $\mathcal{D}_{\text{calib}}$; miscoverage level α; diffusion weights $\lambda_0, \lambda_1, \lambda_2$

Stage 1: Event Prediction Model Training
Train $f(\cdot)$ on historical temporal graphs
Obtain event representations $\mathbf{X}^t = f(\mathcal{G}^{t-h+1:t})$
Compute logits $\hat{\mathbf{y}}^t = F(\mathbf{X}^t)$
Stage 2: Non-Conformity Score Diffusion
for all $(\mathbf{X}^t, \mathbf{y}^t) \in \mathcal{D}_{\text{calib}}$ **do**
  for all $i \in \mathcal{R}$ **do**
    Compute base score $s_i^t \leftarrow |y_i^t - \hat{y}_i^t|$
    Identify topological neighbors $\mathcal{N}_i^{\text{tpo}}$
    Identify temporal neighbors $\mathcal{N}_i^{\text{tpr}}$
    Compute non-conformity scores
  end for
end for
Collect diffused scores $\mathcal{S} = \{s_i^t\}$
Stage 3: Quantile Estimation
Sort $\mathcal{S}$ in ascending order
Compute quantile threshold $\hat{q}$
Stage 4: Efficiency-aware Optimization
Optimize prediction set size by minimizing efficiency loss L_{eff}

C.2 Implementation

We reproduce the event prediction models using publicly available implementations whenever the original authors have provided code. In cases where code is not available, we implement the models from scratch based on the methodological descriptions provided in the corresponding papers. During reproduction, we carefully follow the hyperparameter settings, training procedures, and evaluation protocols specified in the original works to ensure faithful replication of the reported performance. Additionally, any necessary adaptations to accommodate dataset differences or framework updates are explicitly documented to maintain reproducibility. During training, we set the node and edge embedding dimensions to 32 for all event prediction models. Each method is trained for five runs, each consisting of 20 epochs, with a batch size of 2. The reported results are averaged over the five runs. When training the conformal prediction models, we freeze the parameters of the event prediction model. The conformal prediction methods are reproduced based on the implementations provided in the original papers, and their training procedures follow those used for the event prediction models. The transformation function F is defined as a linear function. We perform experiments on Ubuntu 22.04.3 LTS with an NVIDIA A100 and utilize PyTorch to implement all methods.

C.3 Baselines Introduction

We elaborate on the baseline methods considered in our experimental evaluation, with a particular focus on conformal prediction based approaches. Specifically, TPS (Sadinle et al., 2019), APS (Romano et al., 2020), and RAPS (Angelopoulos et al., 2020) represent fundamental conformal prediction techniques, differing primarily in how the target quantile is computed and subsequently used to form the prediction set. These methods serve as canonical references for evaluating coverage guarantees and efficiency trade-offs in conformal inference. DAPS (Zargarbashi et al., 2023) introduces a graph-aware non-conformity score tailored for static graph settings, making it a strong baseline for assessing whether explicitly modeling structural information can improve conformal prediction performance. When applied to event prediction, we extend DAPS to incorporate both topological neighbors and temporal neighbors. CF-GNN (Huang et al., 2023) further advances this line of work by adopting a model-based framework that integrates APS-style non-conformity scores with graph neural networks. Through end-to-end optimization, CF-GNN incorporates topological dependencies directly into the learning process, thereby enhancing the expressiveness of the conformal scoring mechanism under structured data scenarios. Beyond exchangeable settings, NEX (Barber et al., 2023) explicitly targets the violation of the exchangeability assumption in time series data. It addresses temporal distribution shifts by designing non-exchangeable conformal scores, making it particularly relevant for sequential prediction tasks. Similarly, NAPS (Clarkson, 2023) relaxes the exchangeability assumption in graph domains by introducing structure-aware weighting schemes for conformal scores, enabling the method to better reflect relational dependencies in static networks. UGNN (Davis et al., 2024) extends conformal prediction to temporal graphs under the assumption of exchangeability. However, UGNN was origi-

nally designed for static GNN backbones such as GCN, GAT, and GraphSAGE, which restricts its direct compatibility with temporal GNN architectures. To ensure a fair and meaningful comparison, we adapt UGNN by first extracting node representations from the temporal GNN and then applying an additional projection layer to obtain the final predictive outputs.

C.4 Parameter Analysis

To further investigate the impact of different parameters on CFEP, we vary the predefined miscoverage level α and the size of the training data. Experiments are conducted on the Egypt and Iran datasets, and NAPS and NEX are chosen as baseline methods as they exhibit strong performance in the main results. Fig. 5 illustrates that, as the amount of training data decreases, CFEP consistently preserves the predefined coverage while achieving superior efficiency, yielding the best overall performance among all methods. This observation suggests that CFEP can learn stable and robust nonconformity scores even under limited-data regimes by effectively exploiting both temporal and topological neighborhood information. In comparison, DAPS demonstrates better performance than NEX, which is likely attributable to its joint modeling of temporal and topological neighbors, allowing it to remain more resilient to reductions in training data. Fig. 6 shows that, despite the increase in the predefined mis-coverage level, CFEP consistently maintains strong coverage while achieving relatively high efficiency compared with the baseline methods. NEX outperforms DAPS, indicating that jointly incorporating structural and temporal features in DAPS under varying mis-coverage levels may introduce additional noise, which in turn leads to an expansion of the prediction sets.