Appendix A

A.1 Baselines

In this section, we provide comprehensive implementation details for each of the baseline algorithms in our experiments.

PC The PC [Spirtes and Glymour, 1991] algorithm is designed to analyze datasets comprising multiple variables, aiming to uncover conditional independence relationships among variables. PC is a classical and widely-used causal discovery methods, but it is limited to static data and cannot directly handle datasets that include timing information. To address this limitation, we employ a strategy of dividing the original event sequence datasets into time windows. Within each time window, we capture the count of distinct alarms or events present, thus allowing us to employ the PC algorithm. In practice, we find that dividing the dataset into time windows containing 100 events each yields favorable outcomes.

PCMCI The PCMCI [Runge *et al.*, 2019] algorithm integrates principles from partial correlation and Granger causality to estimate causal relationships among variables, encompassing both direct and indirect causal effects. This inclusive approach enables the detection of both linear and non-linear causal interactions. In our experiments, given an event within an event sequence, we define the events occurring within a 5-second time window preceding it as its potential direct cause. Besides, we consider the events within the three preceding time windows as its potential indirect cause.

RL-BIC RL-BIC [Zhu *et al.*, 2019] provides a powerful RL-based framework for addressing the complex task of inferring causal relationships from observational data. By incorporating active exploration and intervention, RL-BIC empowers agents to make informed decisions and effectively uncover causal structures. To apply RL-BIC to our event sequence datasets, we also divide the original datasets into time windows and each time window comprises one hundred events.

ICALiNGAM The ICALiNGAM [Shimizu *et al.*, 2006] algorithm is a powerful approach for estimating causal relationships among variables in a linear non-Gaussian context. It leverages independent component analysis (ICA) and non-Gaussianity to uncover causal structures. However, it is important to note that ICALiNGAM is specifically effective in discovering causal relationships within linear non-Gaussian settings. To successfully apply ICALiNGAM, we adopt a strategy of dividing our datasets into time windows. This partitioning allows us to apply ICALiNGAM to each window independently. In practice, we employ a window size of 100.

ADM4 The ADM4 [Zhou *et al.*, 2013] algorithm offers a convex optimization method for uncovering the underlying network of social influence within a population. It achieves this by modeling event sequence data via the Hawkes process, focusing on the mutual-excitation nature of event dynamics. In our experiments, we configure ADM4 with a decay parameter set to 3, which influences the rate at which historical events' influence decays over time. Additionally, we set the initial coefficients of influence between events between 0.5 and 0.9.

NPHC Non-Parametric Hawkes Cumulant (NPHC) [Achab *et al.*, 2017] is a nonparametric method that supports estimating the matrix of integrated kernels of the multivariate Hawkes process. This method relies on matching the integrated order 2 and order 3 empirical cumulants that represent the simplest set of global observables to recover the Granger causality matrix. In the experiments, we utilize the default parameter settings from the published code which already work well.

CAUSE CAUSE [Zhang *et al.*, 2020] first captures the inherent interdependencies among events via a neural point process model. Then, it leverages an axiomatic attribution method to extract Granger causality statistics from the process. In the experiments, we configure CAUSE with the following specific parameter settings: a batch size of 64, a learning rate of 0.01, and a maximum training epochs of 200.

SHP Structure Hawkes Process (SHP) [Qiao *et al.*, 2023] is an advanced method that leverages the instantaneous effect for learning the Granger causal structure among events types in the discrete-time event sequence. It is featured with the minorization-maximization of the likelihood function and a sparse optimization scheme. In the experiments, we set up SHP with various time interval lengths ranging from 1 to 10. This parameter determines the algorithm's temporal resolution, thereby significantly impacting its performance.

THP Topological Hawkes Process (THP) [Cai *et al.*, 2022] is also a Hawkes process-based causal discovery method. It is a little similar to ADM4. But differently, it proposes to address the problem of causal discovery on non-i.i.d event sequence data via introducing a graph convolution to handle the topological information. It has demonstrated state-of-the-art performance in real-world datasets. In practice, we set up THP with various max hop settings ranging from 0 to 3.

TNPAR Topological Neural Poisson Auto-Regressive Model (TNPAR) [Liu *et al.*, 2024] considers the causal structure as a latent variable and utilizes an amortized inference method to deduce the Granger causal structure among event types. It transforms continuous-time event sequences into discrete-time sequences and leverages a multi-layer perceptron (MLP) to model the event generation process. In the experiments, we set up TNPAR with various time interval lengths ranging from 1 to 10. Similar to that of SHP, this parameter determines the algorithm's temporal resolution. Besides, we also attempt various sparse terms and acyclic penalty coefficients.

A.2 CausalNET

In this paper, all experiments are conducted on a server with NVIDIA Tesla V100 GPUs (32 GB). The server runs on an environment of Ubuntu 18.04 with CUDA 11.8. The parameter settings of CausalNET for the three real-world datasets are shown in Table 1. We adopt the same parameter settings for the Transformer module across these datasets. It consists of 4 blocks. Each block comprises a 4-head self-attention network and a 2-layer feed-forward neural network with 1024 and 512 units. To optimize the causal graph via gradient descent, we employ the Gumbel-Softmax technique with an initial temperature parameter τ of 1, and exponentially anneal it to 0.1 during training. This approach provides greater exploratory capability for the causal graph at the outset of training and gradually reduces the temperature to better balance exploration and exploitation. During training, we utilize the Adam optimizer and set distinct learning rates, η_1 and η_2, for the Transformer module ($f^{(\theta)}$) and the causal graph (i.e., the pa-

rameter matrix ϑ), respectively. In particular, we set the same learning rate for the causal decay matrix (ϕ) as the causal graph. Furthermore, in our opinion, we'd better select the two dataset-related hyper-parameters (i.e., max time lag and max hop) based on both dataset statistics and domain knowledge (e.g., communication delays between devices). However, since domain knowledge for the real-world datasets is currently unavailable, we primarily rely on the dataset statistics to select such hyper-parameters. We have presented the impact of different max time lag and max hop settings on model performance in Figure 3(a) and Figure 3(b) of the main text. The experimental results indicate that even in the worst-case scenario, our proposed CausalNET still outperforms all the baseline models. In addition, our code and data are publicly available at https://github.com/CGCL-codes/CausalNET.

Hyperparameter	Micro-24	Micro-25	IPTV
B	4	4	4
H	4	4	4
L_{ff}	2	2	2
d_k	512	512	512
d_v	512	512	512
d_{ff}	1024	1024	1024
τ	$1 \rightarrow 0.1$	$1 \rightarrow 0.1$	$1 \rightarrow 0.1$
λ_1	10	1	1
λ_2	0.1	0.1	1
λ_3	0.1	0.1	0
η_1	1e-4 $\rightarrow$ 1e-5	1e-4 $\rightarrow$ 1e-5	1e-4 $\rightarrow$ 1e-5
η_2	4e-2 $\rightarrow$ 4e-3	1e-2 $\rightarrow$ 1e-3	1e-2 $\rightarrow$ 1e-3
Batch size	512	512	512
Optimizer	Adam	Adam	Adam
Epochs	200	200	100
ε	0.25	0.25	0.50
ξ	120 s	120 s	24 h
K	1	2	0

Table 1: Hyperparameter Settings

Appendix B

B.1 Qualitative Analysis on IPTV dataset

Since our model is primarily tailored for event sequences with existing topological structures, our main experiments are conducted on two real-world telecommunication network alarm datasets. To explore the applicability of our model in other domains (potentially those without topological structures underlying event sequences), we further test the proposed model on the IPTV dataset [Luo *et al.*, 2015]. Since the causal graph of the IPTV dataset is not typically constrained to be a DAG [Zhang *et al.*, 2020], we remove the DAG-related term from the loss function in this experiment. Figure 1 shows the heatmap for the causal probability graph estimated by the proposed model. The value of the element in the i-th row and j-th column represents the probability of the existence of a causal edge from program i to program j. In other words, it represents the probability that users who watch program i are likely to also watch program j. Overall, the majority of diagonal entries display notably high positive values, suggesting that users tend to watch TV programs within the same category. This really makes sense, as individual users usually engage in frequent viewing of specific types of TV programs that they are interested in. In addition, our model demonstrates

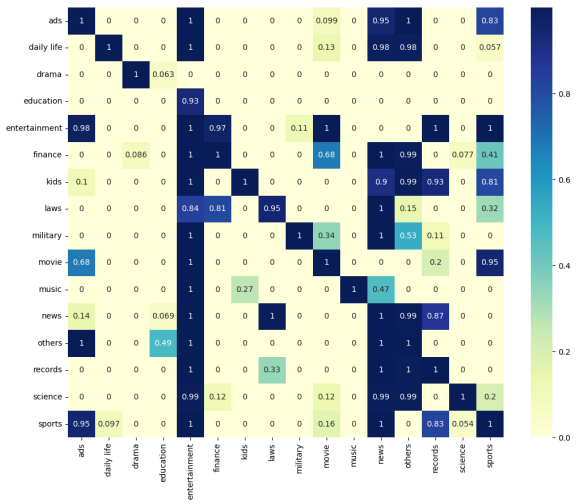


Figure 1: Visualization of the estimated causal probability graph on the IPTV dataset

Dataset	Pruing Strategy	TPR	F1	TPR*	F1*
Micro-24	CausalNET's	0.5839	0.5016	**0.5328**	**0.4883**
	GOLEM's			0.0000	N/A
Micro-25	CausalNET's	0.5270	0.4742	**0.4932**	**0.4679**
	GOLEM's			0.0068	0.0134

Table 2: Comparison of two pruning strategies

some noteworthy phenomena. For example: (1) most programs exhibit a significantly high probability of having edges directed towards "entertainment" and "news". This observation signifies persistent commonalities among users with diverse preferences, indicating their shared interest in watching programs categorized as "entertainment" and "news". (2) Both "Finance" and "military" exhibit edges directed towards "news". This aligns well with our common sense because individuals interested in finance and the military may seek the latest information on these subjects through news reports.

B.2 Comparison of DAG Pruning Strategies

In this section, we conduct a comparison between two DAG pruning algorithms based on real-world datasets. The results are shown in Table 2, with the metrics after pruning marked by "*". As revealed in the table, GOLEM's pruning algorithm yields notably diminished TPR and F1 scores across both datasets. In contrast, our pruning algorithm showcases only a marginal decrease in TPR. As GOLEM conducts pruning by progressively increasing a threshold and removing all edges with weights below the threshold, these results highlight that in the causal structures obtained through training, some non-causal edges have relatively higher weights. This is common, as factors influencing event occurrences encompass not only true causal relationships but also complicated sources of noise. The noise can mislead the causal discovery model to some extent, thereby leading the model to learn causal relationships that do not actually exist and assign relatively high weights to the corresponding edges.

B.3 Attention Score Functions

As mentioned in Methods, we have implemented two different softmax approaches to compute attention weights (or attention scores), which correspond to Equation (4) and Equation (6). While the former allocates attention weights to historical events solely based on the pairwise similarity between the current event and each historical event, the latter introduces additional information about the forthcoming event and meaningless padding events. Here, we evaluate both approaches on two real-world datasets, i.e., Micro-24 and Micro-25. We denote the models using Equation (4) and Equation (6) as CausalNET_D and CausalNET respectively. As shown in Table 3, while both models achieve comparable performance on Micro-24, the latter significantly outperforms the former on Micro-25. This demonstrates the benefits of incorporating the information of the forthcoming event and padding events into the softmax function.

Dataset	Model	F1 ↑	TPR ↑	FPR ↓	AUROC ↑
Micro-24	CausalNET	0.4883	0.5328	0.2027	0.6651
	CausalNET_D	0.4692	0.4453	0.1412	0.6520
Micro-25	CausalNET	0.4679	0.4932	0.1908	0.6512
	CausalNET_D	0.3026	0.2770	0.1719	0.5526

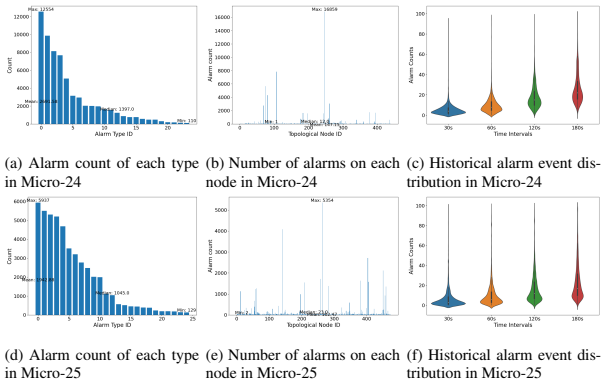
Table 3: Comparison of different softmax approaches

B.4 Detailed Statistics of Datasets

B.4.1 Real-world Datasets

In this section, we present some additional statistics of the two real-world datasets, i.e., 24V_439N_Microwave (Micro-24) and 25V_474N_Microwave (Micro-25), which record alarm events in real-world telecommunication networks. Although both datasets are collected from devices in telecommunication networks, they are independent of each other. Specifically, they differ in terms of alarm types, devices, and event sequence lengths. For instance, the same alarm ID in the two datasets may represent different actual alarm event types. As shown in Figure 2, the event distributions within both datasets are highly uneven. On the one hand, there are significant disparities in event occurrence frequencies across different types. For example, in the first dataset, the most frequent event (type: 0) occurred 12,554 times, while the least frequent event (type: 23) occurred only 110 times. On the other hand, event occurrences across numerous topological nodes vary greatly, ranging from thousands to merely one or two occurrences. This uneven event distribution presents challenges for causal discovery tasks. In practice on such two datasets, we find that CausalNET performs the best when removing the topological node embedding in Equation (2) and the topological node prediction loss in Equation (13). This might be related to the significantly unbalanced distribution of the events across topological nodes. In future work, we will delve deeper into how the distribution of events impacts the performance of causal discovery models.

In addition, we have tabulated the historical event counts within the preceding 30s, 60s, 120s, and 180s of each event. The results are shown in Figure 2(c) and Figure 2(f). For clarity, we have truncated the data where the historical event count exceeds 100. The distribution of historical events differs between the two datasets, affirming the need to set distinct max time lag parameters (ξ) for different datasets.



(a) Alarm count of each type in Micro-24 (b) Number of alarms on each node in Micro-24 (c) Historical alarm event distribution in Micro-24

(d) Alarm count of each type in Micro-25 (e) Number of alarms on each node in Micro-25 (f) Historical alarm event distribution in Micro-25

Figure 2: Additional statistics of two real-world datastes

B.4.2 Synthetic Datasets

In addition to the aforementioned real-world datasets, we also generate a range of synthetic datasets via gcatle's API[1], which simulates event sequences based on the topological Hawkes process [Cai *et al.*, 2022] engineered with a classical exponential decay kernel function. In practice, to generate datasets with different event interactions and temporal effects, we reformulate the original topological Hawkes process by replacing the default kernel function with some other kernel functions from the field of point process and event modeling. Specifically, we introduce the Weibull distribution [Rinne, 2008], whose probability density function is defined as:

$$f(x;k,\lambda) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (1)$$

where k is the shape parameter, and λ is the scale parameter. By adjusting the shape parameter k, it can be tailored to fit various shapes of data distributions that correspond to different kernel functions. In particular, when the shape parameter k is equal to 1, the Weibull distribution degenerates into the exponential distribution, which corresponds to the classical exponential decay kernel function. Finally, the distinctions among our synthetic datasets lie in the kernel function, the number of event types, the number of topological nodes, and the event sequence length (total events). Without further specification, each synthetic dataset utilizes an exponential decay function as the kernel, comprising 30 event types, 60 topological nodes, and an event sequence length of 30,000.

References

[Achab *et al.*, 2017] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1–10. PMLR, 2017.

[1]https://github.com/huawei-noah/trustworthyAI/tree/master/gcastle

[Cai *et al.*, 2022] Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. Thps: Topological hawkes processes for learning causal structure on event sequences. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Liu *et al.*, 2024] Yuequn Liu, Ruichu Cai, Wei Chen, Jie Qiao, Yuguang Yan, Zijian Li, Keli Zhang, and Zhifeng Hao. Tnpar: Topological neural poisson auto-regressive model for learning granger causal structure from event sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 20491–20499, 2024.

[Luo *et al.*, 2015] Dixin Luo, Hongteng Xu, Yi Zhen, Xia Ning, Hongyuan Zha, Xiaokang Yang, and Wenjun Zhang. Multi-task multi-dimensional hawkes processes for modeling event sequences. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3685–3691, 2015.

[Qiao *et al.*, 2023] Jie Qiao, Ruichu Cai, Siyu Wu, Yu Xiang, Keli Zhang, and Zhifeng Hao. Structural hawkes processes for learning causal structure from discrete-time event sequences. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 5702–5710, 2023.

[Rinne, 2008] Horst Rinne. *The Weibull distribution: a handbook*. CRC press, 2008.

[Runge *et al.*, 2019] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019.

[Shimizu *et al.*, 2006] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[Spirtes and Glymour, 1991] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

[Zhang *et al.*, 2020] Wei Zhang, Thomas Panum, Somesh Jha, Prasad Chalasani, and David Page. Cause: Learning granger causality from event sequences using attribution methods. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11235–11245. PMLR, 2020.

[Zhou *et al.*, 2013] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 641–649. PMLR, 2013.

[Zhu *et al.*, 2019] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019.