METP: Multi-Granularity Integration of External Covariates for Temporal Point Processes

Boyang Li[1*], **Lingzheng Zhang**[2*], **Fugee Tsung**[3†], **Xi Zhang**[1†]
[1]Peking University, Beijing, China
[2]The Hong Kong University of Science and Technology (GZ), Guangzhou, China
[3]The Hong Kong University of Science and Technology, Hong Kong, China
2101112018@stu.pku.edu.cn, lingzhengzhang01@gmail.com, season@ust.hk, xi.zhang@pku.edu.cn

Abstract

Accurate modeling of temporal point processes is critical for reliable event forecasting and informed decision-making. While historical event sequences provide a foundation for intensity estimation, existing approaches often neglect external covariates whose lagged effects impact future intensities across multiple temporal granularities. To address this gap, we propose Multi-Granularity Integration of External Covariates for Temporal Point Processes (METP), a framework for incorporating lagged external influences into intensity modeling. METP extracts periodic structures and decomposes external covariate series into multiple temporal granularities. At each granularity, a lag-aware calibration module is introduced to align covariates with event dynamics. Finally, a hierarchical mixture-of-experts strategy is employed to integrate the multi-granular external covariates with historical event embeddings, enabling a representation of the conditional intensity function with enhanced information. Extensive experiments on public and proprietary datasets demonstrate that METP consistently outperforms existing methods in predictive accuracy.

Extended version —
https://github.com/lbylbylby123456/AAAI-METP

Introduction

Temporal Point Processes (TPPs) have emerged as a powerful mathematical framework for modeling stochastic event sequences in continuous time (Gu 2021). Their versatility is demonstrated by successful applications across diverse domains, including e-commerce systems (Boyd et al. 2020), social network analysis (Karpukhin, Shipilov, and Savchenko 2024), and clinical informatics (Xiao et al. 2025). The integration of deep learning methodologies has significantly advanced TPP modeling, with neural network-based approaches offering enhanced capabilities for capturing complex temporal patterns (Shou et al. 2023). By leveraging the nonlinear approximation capacity of neural architectures, these models capture complex temporal dependencies from historical events, enabling more accurate prediction of future event intensities (Gracious and Dukkipati 2025).

Despite the promising performance of existing methods in event prediction, many TPP models pay insufficient attention to external temporal covariates (Zuo et al. 2020; Zhang et al. 2020; Meng et al. 2024) and neglect their lagged effects on event intensities (Shams Eddin and Gall 2024), as illustrated in Fig. 1(b). Given that these covariates frequently correlate with events, neglecting them can lead to missed key dependencies, resulting in inaccurate intensity estimation and biased event time prediction. Incorporating external variables is therefore essential for enhancing predictive accuracy. Moreover, external temporal variables sometimes exhibit periodic patterns, which form a fundamental aspect of temporal modeling (Li et al. 2021a) and play a critical role in diverse applications such as consumption forecasting (Lin et al. 2024), healthcare prediction (Xiao et al. 2025), and industrial prognostics (Chen et al. 2025). To capture these periodicities, existing methods typically employ Fourier transforms to extract dominant frequency components (Mu, Shahzad, and Zhu 2025), analyzing subsequences of the original series for subsequent modeling tasks. However, relying solely on the original subsequences often results in limited expressiveness, not accounting for the necessity to model both long-term trends and short-term fluctuations, which are essential factors influencing event intensities. This limitation highlights the need to effectively model periodic patterns across multiple temporal granularities and to integrate their combined influences.

Furthermore, although some studies attempt to analyze the influence of external variables, these models often do not account for the lagged effects of external series (Zhang et al. 2024a). In real-world scenarios, such factors typically evolve asynchronously with event sequences, leading to temporal misalignments that challenge conventional modeling frameworks (Zhang et al. 2023; Kuang et al. 2024). Recent methods typically employ the time point within a fixed-length period before the event to approximate such lag effects (Shams Eddin and Gall 2024). However, they often neglect the decay influence of historical time series at different time points on current events. In addition, the absence of mechanisms for capturing periodic structures and multi-granularity temporal variations restricts their capacity to represent the dynamic heterogeneity of external influ-

[*]These authors contributed equally.
[†]Corresponding authors.

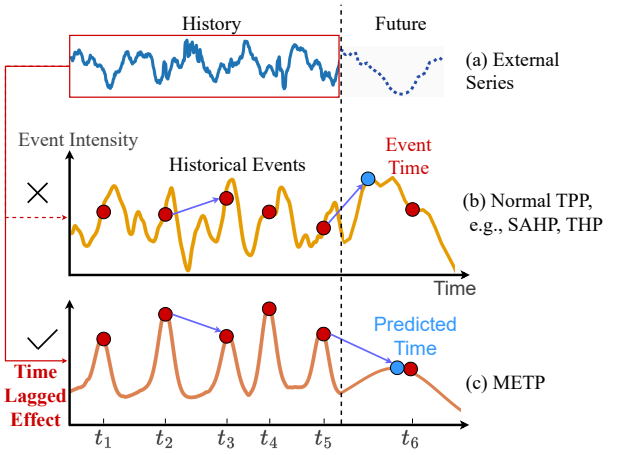


Figure 1: Comparison of Normal TPP and METP in Event Time Prediction with Lagged Effects.

ences. Compounding this issue is the unavailability of future environmental time series, making it difficult to align historical covariates with long-term event forecasting. These challenges highlight the need for advanced models that can dynamically synchronize historical external series with event processes while capturing their multi-granular dependencies to improve prediction accuracy.

To address these limitations, we propose a novel framework, Multi-Granularity Integration of External Covariates in Temporal Point Processes (METP), which incorporates external series more effectively than standard approaches, as illustrated in Fig. 1(c). The method integrates periodic patterns from external sequences across multiple temporal granularities. By aligning the weights of a causal self-attention mechanism with a prior reversed geometric distribution, it adaptively models the lagged effects of external covariates at each granularity. Finally, a hierarchical mixture-of-experts (MoE) framework integrates the multi-granular external covariates with historical event sequence embeddings, leading to significant improvements in event intensity forecasting accuracy. The key contributions of this work are summarized as follows:

- We propose a multi-granularity, periodicity-aware attention architecture. It models the lagged effects of external covariates on event occurrence by aligning the distributions of reversed geometric priors and causal attention scores. To the best of our knowledge, this is the first method that enables adaptive alignment of lagged effects in temporal point processes.
- We integrate historical external covariates through a hierarchical granular mixture-of-experts strategy, combined with historical event embeddings to predict future event intensities.
- Experiments on a proprietary dataset and three public benchmarks demonstrate consistent improvements, validating the robustness and generalizability of the proposed framework.

Related Work

Temporal Point Process Methods

Temporal Point Processes (TPP) provide a probabilistic framework for modeling discrete event sequences in continuous time. Traditional parametric models, such as the Hawkes process (Hawkes 1971), employ parametric and stationary intensity functions, which limit their ability to capture complex and dynamic temporal dependencies. To overcome these limitations, recent approaches have introduced neural TPP models that leverage data-driven architectures to capture intricate temporal dependencies. RMTPP (Du et al. 2016) integrates recurrent neural networks with TPP to learn irregular temporal structures. IFTPP (Shchur, Biloš, and Günnemann 2019) eliminates the need for explicit intensity functions by using normalizing flows to directly model inter-event time distributions.

With the rise of attention-based mechanisms, THP (Zuo et al. 2020) introduces a Transformer-based self-attention structure to capture both short- and long-range dependencies. FTHP (Isik et al. 2023) further replaces standard attention with learnable triggering kernels and gating functions to enhance modeling effectiveness and interpretability. To address the distinct characteristics of short- and long-term prediction tasks, HoTPP (Karpukhin, Shipilov, and Savchenko 2024) introduces a long-horizon forecasting benchmark and proposes a new evaluation metric (T-mAP). ITHP (Meng et al. 2024) extends Transformer Hawkes models with nonlinear inter-type interaction modeling for improved interpretability. TPP-Gaze (D'Amelio et al. 2025) incorporates spatial-temporal attention for modeling visual attention dynamics in eye movement sequences. Despite these advancements in modeling internal event dependencies, challenges remain in effectively leveraging external covariates. In practice, these external variables often exhibit periodic behaviors (Lin et al. 2024). Their influences on event dynamics manifest through both long-term trends and short-term perturbations across multiple temporal granularities (Li et al. 2021a). They also tend to produce lagged and temporally misaligned effects, which are difficult to capture using static lag structures.

Most existing models do not explicitly model complex lagged effects, which lead to an incomplete representation of the dynamic heterogeneity and multi-granular temporal structures commonly observed in real-world scenarios. As a result, the absence of effective mechanisms for incorporating external information continues to limit the predictive accuracy and generalizability of current TPP frameworks.

Misalignment between Covariate and Event

Temporal misalignment is a fundamental challenge in time series modeling, where outputs are influenced not only by concurrent inputs but also by temporally lagged signals. Recent studies have proposed solutions such as hierarchical latent space decomposition (Li et al. 2021b), attention-based inter-feature delay modeling (Dai et al. 2024), and component-wise attention normalization (Deng et al. 2024). For irregular time series, asynchronous graph diffusion (Zhang et al. 2024c) and transformable time-aware convo-

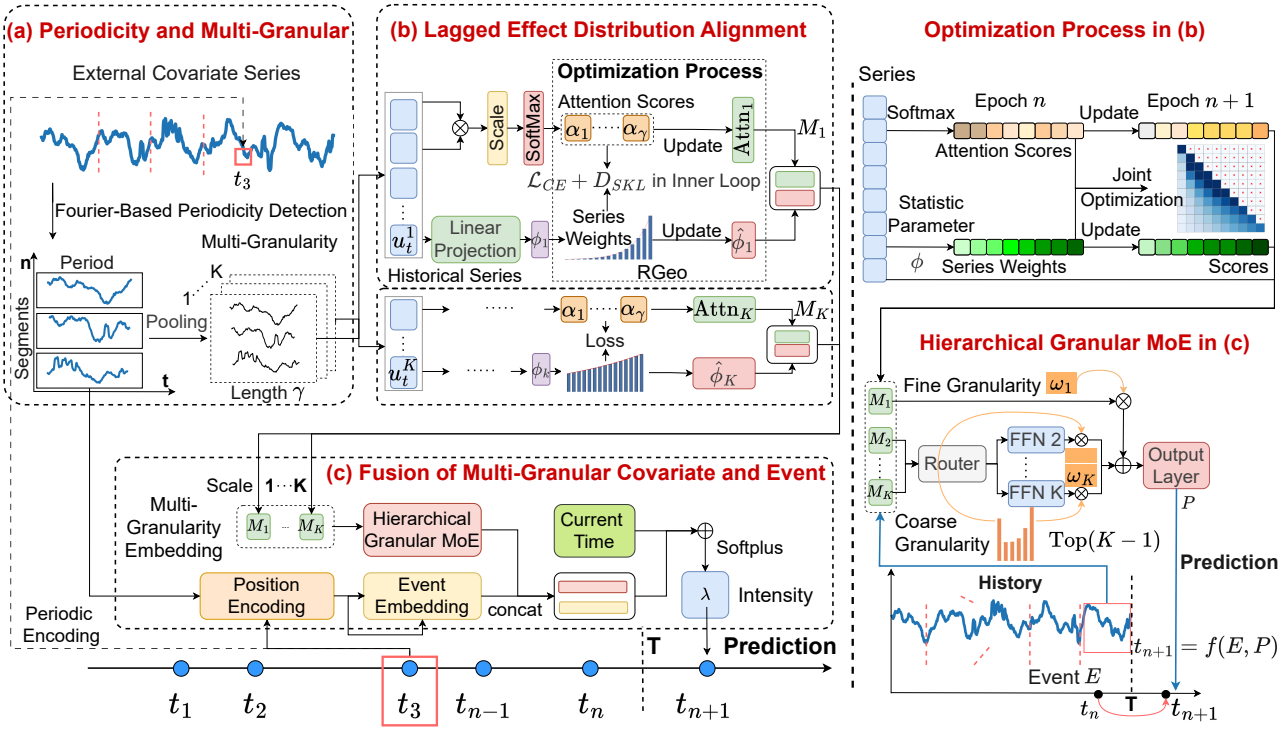


Figure 2: The framework of METP.

lutional networks have been introduced to achieve spatial-temporal alignment. These methods have demonstrated effectiveness in continuous, high-density settings.

In contrast, sparse event-driven scenarios, such as those encountered in temporal point processes, present unique challenges. Temporal misalignment between event sequences and external variables often arises due to lagged effects and indirect interactions. Existing approaches commonly adopt fixed-length lag windows (Shams Eddin and Gall 2024), which do not explicitly account for variable periodic structures and multi-granularity dependencies, limiting adaptability to dynamic external influences.

Recent developments, including diffusion-based alignment (Gao, Cao, and Chen 2025) and cross-modal matching frameworks (Liu et al. 2025), provide improved alignment mechanisms. However, they generally neglect the incorporation of prior knowledge regarding lag structures and the dynamic adjustment of alignment strategies across multiple temporal granularities, which could result in inaccurate predictions when modeling complex temporal dependencies. These limitations highlight the need for alignment strategies that are both prior-informed and granularity-aware to effectively capture the temporal misalignment in sparse, event-based data.

Methodology

Problem Formulation

Let $\mathcal{S}_{1:n} = \{t_1, t_2, \ldots, t_n\} \in \mathbb{R}^n$ denote a sequence of n observed event timestamps, where t_j represents the timestamp of the j-th event. The goal is to predict the time of the next event, t_{n+1}. External variables observed at discrete time steps $t \in [1, T]$ are denoted by x_t, which are assumed to influence the temporal dynamics of event occurrences. In contrast to traditional methods that focus exclusively on past event histories, the proposed framework incorporates external inputs to capture their impact on events. Given the historical event sequence $\mathcal{S}_{1:n}$ and the corresponding external input sequence $\mathcal{X}_T = \{x_1, \ldots, x_T\}$, the conditional intensity function $\lambda(t)$ at time $t > T$ is defined as

$$\lambda(t) = \mathcal{F}\left(t > T | \mathcal{S}_{1:n}, \mathcal{X}_T\right), \tag{1}$$

where $\mathcal{F}(\cdot)$ jointly models temporal dependencies and dynamic external effects, enabling improved next-event time prediction under non-stationary conditions. Following (Zuo et al. 2020), the conditional density of the next event time given the known observations is defined based on the conditional intensity function $\lambda(t)$ as

$$\kappa(t) = \lambda(t) \exp\left(-\int_{t_n}^{t} \lambda(\tau)\, d\tau\right), \tag{2}$$

where $t \geq t_n$ and t_n denotes the timestamp of the most recent observed event. Based on this density, the expected arrival time of the next event after T is obtained via the conditional expectation:

$$t_{n+1} = \int_{T}^{\infty} t \cdot \kappa(t)\, dt. \tag{3}$$

This formulation forms the theoretical basis for our predictive framework, which aims to jointly model temporal de-

pendencies and exogenous environmental influences for accurate event time forecasting.

Framework Overview. The proposed METP framework integrates three key modules: periodicity and multi-granular encoding, alignment of external lag effects, and fusion of covariates and events for event intensity prediction, as illustrated in Fig. 2.

Periodicity and Multi-Granular Encoding

To effectively model temporal dependencies and capture periodic patterns in the external time series, we extract the dominant periodicity from the external time series $\{x_1, x_2, \ldots, x_t\}$ using the Discrete Fourier Transform. The frequency with the highest magnitude, excluding zero frequency, is identified to determine the dominant period δ. The detailed algorithm is provided in the Appendix. A sliding window then extracts the most recent segment of length δ:

$$x_{t-\delta+1:t} = [x_{t-\delta+1}, \ldots, x_t]. \tag{4}$$

To incorporate multiple past cycles, we construct a compact periodic time matrix $\mathbf{X}(t) \in \mathbb{R}^{\delta \times m}$, where δ is the period length and m is the number of past periods. Each row corresponds to a complete historical period, and each column represents the same relative position across periods:

$$\mathbf{X}(t) = \left[x_{t-\delta+1:t}, \ldots, x_{t-m\delta+1:t-(m-1)\delta}\right]^\top. \tag{5}$$

Right-padding is applied when insufficient data points are available. To capture both long-term trends and short-term fluctuations, a hierarchical pooling strategy is applied to $\mathbf{X}(t)$ to analyze information across different granularities. We define a set of granularities $\{s_k\}_{k=1}^K$, where each s_k represents the length of a non-overlapping pooling window. For each granularity, a max pooling is performed:

$$\mathbf{U}_k(t) = \text{MaxPool}_{s_k}\left(\mathbf{X}(t)\right) \in \mathbb{R}^{\left\lfloor \frac{\delta}{s_k} \times m \right\rfloor}, \tag{6}$$

where $\text{MaxPool}_{s_k}(\cdot)$ denotes a temporal pooling operator with kernel and stride size s_k. This process facilitates the extraction of both short-term and long-term temporal dependencies. Subsequently, the temporal positioning of events is encoded to distinguish events based on their absolute occurrence times. We adopt a sinusoidal position encoding scheme. While prior work such as (Xiao et al. 2025) determines periodicity based on sampling patterns from the event sequence, our approach differentiates itself by directly deriving the encoding period δ from the dominant periodicity present in the external covariates. This enables a more covariate-aware temporal encoding. Specifically, each timestamp t is mapped to a deterministic vector $e(t) \in \mathbb{R}^d$:

$$\begin{aligned} e(t) = \sqrt{\tfrac{2}{d}}\Big[&\sin\!\left(\tfrac{2\pi\omega_1 t}{\delta}\right),\ \cos\!\left(\tfrac{2\pi\omega_1 t}{\delta}\right),\ \ldots, \\ &\sin\!\left(\tfrac{2\pi\omega_{d/2} t}{\delta}\right),\ \cos\!\left(\tfrac{2\pi\omega_{d/2} t}{\delta}\right)\Big], \end{aligned} \tag{7}$$

where $\{\omega_i\}_{i=1}^{d/2}$ are frequency components defined as $\omega_i = 10000^{\frac{2(i-1)}{d}}$. This design introduces a data-driven temporal basis with external dynamics, improving the model sensitivity to periodic structures in real-world event sequences.

Alignment of Lagged Effect Distributions

Given the discrete time steps and strong temporal dependencies between adjacent inputs, a non-symmetric decay modeled by a reversed geometric distribution offers an effective prior for capturing lagged effects. Flexibility is further enhanced by incorporating an additional causal self-attention mechanism. This mechanism captures time-aware dependencies between different historical moments and the current event time, with attention scores adaptively reflecting their relative importance. To conform to established temporal-decay patterns while retaining data-driven flexibility, reversed-geometric weights are precisely aligned with causal self-attention scores, yielding an accurate portrayal of lagged effects. For simplicity, the time index t is omitted in the notation. Let $\mathbf{u}_k$ denote the vector corresponding to the latest time in the temporal sequence $\mathbf{U}_k$. The decay parameter is given by $\phi_k = \mathbf{u}_k \mathbf{w}_k$, where w_k is a learnable projection vector. The corresponding normalized reversed geometric prior distribution models the temporal influence as

$$r_k(j) = \frac{(1-\phi_k)^{j-1} \cdot \phi_k}{1-(1-\phi_k)^{\gamma_k}}, \tag{8}$$

where $j = 1, 2, \ldots, \gamma_k$, and γ_k denotes the maximum delay window at the k-th granularity. To flexibly learn the relevance distribution in a data-driven manner (Xu et al. 2021), we adopt a time-aware causal self-attention mechanism. For each granularity k, the query and key vectors are computed as $q_k = \mathbf{u}_k \mathbf{w}_1$, $k_k = \mathbf{u}_k \mathbf{w}_2$, where $\mathbf{w}_1$ and $\mathbf{w}_2$ are learnable projection matrices. To capture the influence of all key vectors on the query vector at granularity k, the raw attention scores are computed as

$$s_k^j = \frac{q_k^\top k_k^j}{\sqrt{l}}, \tag{9}$$

where k_k^j represents the j-th key vector, and l is the dimensionality of the vectors. For the γ_k-th query position, the attention score α_k^j assigned to the j-th key is given by the softmax function as

$$p_k(j) = \alpha_k^j = \frac{\exp(s_k^j)}{\sum_{i=1}^{\gamma_k} \exp(s_k^i)}, \tag{10}$$

where the distribution $p_k(j)$ quantifies the attention score at the j-th time point when predicting the current query at granularity k. To align the learned reverse geometric distribution r_k with the attention score distribution p_k, we adopt the Symmetric Kullback–Leibler (SKL) divergence:

$$\mathrm{D}_{\mathrm{SKL}}(k) = \mathrm{D}_{\mathrm{KL}}(p_k \,\|\, r_k) + \mathrm{D}_{\mathrm{KL}}(r_k \,\|\, p_k), \tag{11}$$

where

$$\mathrm{D}_{\mathrm{KL}}(p_k \,\|\, r_k) = \sum_{j=1}^{\gamma_k} p_k(j) \log \frac{p_k(j)}{r_k(j) + \epsilon}, \tag{12}$$

where $\epsilon > 0$ is a small constant to ensure numerical stability. Subsequently, the aggregated hidden representation at time t for granularity k is computed as the attention-weighted sum of historical hierarchical vectors:

$$z_t^k = \sum_{j=1}^{\gamma_k} \alpha_k^j u_k^{t-j+1}, \tag{13}$$

where u_k^{t-j+1} denotes the hierarchical representation at granularity k and delay step j. The prediction $\hat{y}_t$ of future event occurrence is then obtained via a linear transformation and sigmoid activation:

$$\hat{y}_t^k = \sigma(w_z z_t^k + b_z), \tag{14}$$

where $w_z, b_z \in \mathbb{R}$, and $\sigma(\cdot)$ is the sigmoid function. The loss function is minimized to optimize model parameters:

$$\mathcal{L}_k = \sum_{t=1}^{T} \mathcal{L}_{CE}(y_t, \hat{y}_t^k) + \eta \, \mathrm{D}_{\mathrm{SKL}}(k), \tag{15}$$

where $\mathcal{L}_{CE}(\cdot,\cdot)$ is the weighted cross-entropy with a weighting factor determined by the event ratio. $y_t \in \{0,1\}$ indicates whether an event occurs at time t, and $\eta > 0$ is a trade-off parameter that balances predictive accuracy and regularization. The cross-entropy term drives prediction accuracy, while the SKL term constrains the attention weight distribution to remain consistent with the prior, thereby regularizing temporal structure learning.

Fusion of Multi-Granular Covariates and Events

To model the conditional intensity function of the event e_{j+1} occurring at time t, given the j-th event at time t_j, we fuse aligned external variables with historical event embeddings to capture both past events and external influences on future occurrences. The aligned external representation at granularity k is defined as:

$$\mathbf{P}_k(t) = \frac{1}{Z_k} \sum_{i=1}^{\gamma_k} \alpha_k^i \cdot d_k(t - t_j) \cdot \mathbf{u}_k^{t_j - i + 1}, \tag{16}$$

where the attention weights α_k^i are derived from causal self-attention at granularity k, and $\mathbf{u}_k^{(t_j - i + 1)}$ denotes the external covariate embedding at the i-th lag step with respect to event time t_j. The normalization constant is defined as $Z_k = \sum_{i=1}^{\gamma_k} \alpha_k^i$. The scalar lagged weight is defined as $d_k(t - t_j) = (1 - \phi_k)^{t - t_j}$, which applies an exponential decay based on the temporal distance between the current time t and t_j. This method separates the effects of attention-based relevance and temporal decay, enabling more interpretable modeling of lagged external influences. Multi-granularity embeddings $\{\mathbf{P}_k\}_{k=1}^K$ are integrated using a hierarchical granular MoE mechanism. To ensure the analysis remains grounded in the original temporal resolution, the embedding at the first granularity is always included. The remaining $K - 1$ expert embeddings are dynamically selected at each time step t, yielding the aggregated representation:

$$\mathbf{P}(t) = \pi_1 \mathbf{P}_1(t) + \sum_{k \in \mathcal{K}_t} \pi_k \mathbf{P}_k(t), \tag{17}$$

where the set $\mathcal{K}_t \subseteq \{2, \ldots, K\}$ denotes the indices of the top-$(K-1)$ selected experts at time t, and $\boldsymbol{\pi} = (\pi_1, \pi_k)_{k \in \mathcal{K}_t}$ are the learned mixture coefficients, producing a unified environment-aware context. The historical event embedding sequence $\mathbf{E} = [e_1, \ldots, e_n]$ is processed through multi-head self-attention and feed-forward layers:

$$\mathbf{M} = \mathrm{MultiHeadAttention}(\mathbf{E}), \tag{18}$$

$$\mathbf{H} = \mathrm{LayerNorm}(\mathbf{M} + \mathrm{FFN}(\mathbf{M})). \tag{19}$$

The conditional intensity function at time t is formulated as

$$\lambda(t|\mathbf{H}_t) = \mathrm{softplus}\left(\rho \frac{t - t_i}{t_i} + \mathbf{w}_{\mathrm{a}}^\top \mathbf{h}_{\mathrm{a}}(t) + b\right), \tag{20}$$

where

$$\mathbf{w}_{\mathrm{a}} = [\mathbf{w}_h, \mathbf{w}_e], \quad \mathbf{h}_{\mathrm{a}}(t) = [\mathbf{h}_t, \mathbf{p}(t)]. \tag{21}$$

Here, ρ controls the temporal decay, $\mathbf{w}_h$ and $\mathbf{w}_e$ are learnable weight vectors, and b is a bias term. The softplus activation function guarantees the non-negativity of $\lambda(t|\mathbf{H}_t)$. This formulation enables modeling of intensity by jointly capturing dependencies in historical events and covariates.

Training Objective

The negative log-likelihood loss $\mathcal{L}_p$ is constructed based on the conditional intensity function and approximated by a discrete summation:

$$\mathcal{L}_p = -\sum_{i=1}^{n} \log \lambda(t_i|\mathcal{H}_{t_i}) + \sum_{j=1}^{T} \lambda(j|\mathcal{H}_j), \tag{22}$$

where $\lambda(t|\mathcal{H}_t)$ denotes the conditional intensity at time t, and the integral is estimated over T discrete time steps. The overall loss function is given by:

$$\mathcal{L} = \mathcal{L}_p + \sum_{k=1}^{K} \mathcal{L}_k, \tag{23}$$

where $\mathcal{L}_k$ denotes the alignment loss at granularity level k. A two-stage optimization is adopted for stability. Details of the overall algorithm and training procedure are provided in the Appendix. Additionally, based on the learned intensity function, the expected time of the next event in Equation (3) is approximated by:

$$\hat{t}_{j+1} = \sum_{t=T}^{M} t \cdot p(t|\kappa(t)), \tag{24}$$

where M is the number of discrete time points considered. This approximation enables practical prediction of event timings in discrete settings.

Experiment

Experiment Setting

Datasets. We evaluate our method on one proprietary dataset and three publicly available benchmarks. The proprietary **Gasoline Transaction Dataset (GTD)** records longitudinal gasoline transaction data, with external covariates representing sales prices. The **Tianchi-Walmart Storm Sales Dataset**[1] includes product-level sales data across multiple stores, accompanied by external temperature time series. The **Elevator Fault Dataset**[2] involves event modeling for annotated elevator fault types, with associated indi-

[1] https://tianchi.aliyun.com/dataset/89813

[2] https://www.kaggle.com/datasets/ziya07/elevator-fault-monitoring-and-early-warning-system

Model	GTD			Tianchi			Fault			Earthquake			Avg. NLL	Avg. NMAE	Avg. NRMSE	Avg. Rank
	NLL	NMAE	NRMSE	NLL	NMAE	NRMSE	NLL	NMAE	NRMSE	NLL	NMAE	NRMSE				
RMTPP (2016)	$15.90_{0.24}$	$0.94_{0.02}$	$1.86_{0.05}$	$1.99_{0.02}$	$\mathbf{0.58}_{0.03}$	$4.58_{0.03}$	$1.70_{0.04}$	$0.42_{0.01}$	$\mathbf{0.77}_{0.01}$	$12.01_{0.28}$	$0.94_{0.02}$	$9.32_{0.10}$	7.90	0.72	4.13	7.13
IFTPP (2019)	$14.85_{0.18}$	$0.97_{0.03}$	$1.79_{0.05}$	$2.56_{0.10}$	$1.50_{0.03}$	$6.01_{0.06}$	$2.06_{0.03}$	$0.63_{0.01}$	$1.05_{0.03}$	$12.73_{0.19}$	$2.19_{0.06}$	$10.05_{0.09}$	8.55	1.32	4.73	7.38
THP (2020)	$9.81_{0.07}$	$0.49_{0.01}$	$1.08_{0.02}$	$1.60_{0.04}$	$1.25_{0.05}$	$3.87_{0.03}$	$1.50_{0.04}$	$0.36_{0.02}$	$0.98_{0.03}$	$11.24_{0.12}$	$1.12_{0.02}$	$9.02_{0.03}$	6.54	0.81	3.23	3.63
SAHP (2020)	$14.56_{0.10}$	$0.78_{0.04}$	$1.49_{0.03}$	$2.56_{0.10}$	$1.50_{0.03}$	$6.01_{0.06}$	$2.84_{0.05}$	$0.71_{0.02}$	$1.12_{0.04}$	$13.24_{0.15}$	$2.31_{0.03}$	$10.01_{0.04}$	8.80	1.33	4.66	6.88
A-NHP (2022)	$9.97_{0.08}$	$0.62_{0.03}$	$1.64_{0.04}$	$1.76_{0.05}$	$0.46_{0.05}$	$5.61_{0.03}$	$3.57_{0.08}$	$1.04_{0.03}$	$3.82_{0.05}$	$18.26_{0.10}$	$0.99_{0.02}$	$11.28_{0.06}$	8.89	0.78	5.09	5.75
ITHP (2024)	$9.09_{0.09}$	$0.40_{0.02}$	$1.01_{0.04}$	$1.64_{0.08}$	$0.98_{0.04}$	$3.19_{0.02}$	$1.31_{0.05}$	$0.41_{0.03}$	$0.99_{0.05}$	$10.86_{0.04}$	$0.91_{0.10}$	$\mathbf{8.77}_{0.03}$	5.72	0.68	3.24	2.63
XTSFormer (2025)	$10.15_{0.32}$	$1.23_{0.04}$	$1.98_{0.03}$	$1.60_{0.04}$	$1.41_{0.03}$	$3.45_{0.05}$	$1.50_{0.04}$	$0.38_{0.02}$	$0.95_{0.03}$	$16.58_{0.01}$	$0.91_{0.04}$	$10.75_{0.08}$	7.46	0.98	4.28	5.88
METP	$\mathbf{8.25}_{0.08}$	$\mathbf{0.38}_{0.03}$	$\mathbf{0.90}_{0.02}$	$\mathbf{1.55}_{0.02}$	$0.98_{0.02}$	$\mathbf{3.15}_{0.02}$	$\mathbf{1.23}_{0.04}$	$\mathbf{0.36}_{0.02}$	$0.84_{0.03}$	$\mathbf{10.06}_{0.08}$	$\mathbf{0.88}_{0.02}$	$8.91_{0.07}$	**5.27**	**0.65**	**3.20**	**1.63**

Table 1: Quantitative performance comparison across GTD, Tianchi, Fault, and Earthquake datasets using NLL, NMAE, and NRMSE. Average scores over all available datasets are reported on the right. Lower values indicate better performance. The last column shows the average ranking of each model over all metrics.

Model Variant	GTD			Tianchi		
	NLL	NMAE	NRMSE	NLL	NMAE	NRMSE
METP (Full Model)	**8.25**	**0.38**	**0.90**	**1.55**	**0.98**	**3.15**
w/o External Variables	9.56	0.49	1.12	1.92	1.56	3.82
w/o Lagged Weights	8.96	0.44	1.03	1.85	1.25	3.41
w/o Multi-Scale Structure	8.31	0.42	0.94	1.62	1.07	3.32
w/o Periodic Structure	8.65	0.43	1.15	1.70	1.09	3.30

Table 2: Ablation study results on the GTD dataset and the public benchmark dataset (Tianchi).

cator time series as external features. The **Global Earthquake Dataset**[3] contains large-scale geophysical earthquake records, with extreme weather conditions as external variables. Detailed descriptions are provided in the Appendix.

Baselines and Evaluation. We compared our method with several state-of-the-art TPP baselines. **RMTPP** (Du et al. 2016) combines RNNs with TPPs to capture temporal dependencies. **IFTPP** (Shchur, Biloš, and Günnemann 2019) uses normalizing flows to model inter-event times without explicit intensity functions. **THP** (Zuo et al. 2020) employs Transformer self-attention for short- and long-term dependencies. **SAHP** (Zhang et al. 2020) replaces RNNs with attention and encodes inter-event times via sinusoidal phase shifts. **A-NHP** (Yang, Mei, and Eisner 2022) uses symbolic time-varying embeddings to improve parallelizability. **ITHP** (Meng et al. 2024) models nonlinear inter-type dependencies for better interpretability. **NJDTPP** (Zhang et al. 2024b) defines intensities via neural jump-diffusion SDEs for unified dynamic modeling. **XTSFormer** (Xiao et al. 2025) captures multi-scale irregular clinical events with cross-temporal-scale Transformers.

To evaluate the model effectiveness in predicting event times, we employ three widely used metrics: Negative Log-Likelihood (NLL), Normalized Mean Absolute Error (NMAE), and Normalized Root Mean Square Error (NRMSE). NLL evaluates the probabilistic fit between the predicted intensity function and the observed event sequence, while NMAE and NRMSE assess the normalized deviations between predicted and actual time, with lower values across all metrics indicating better performance.

Implementation Details. The proposed model is implemented in PyTorch using a 3-layer Transformer encoder architecture with 4 attention heads. The model is trained with the Adam optimizer (learning rate = $1e{-}3$, batch size = 16) for 100 epochs. For numerical stability in divergence estimation, we apply a small constant $\epsilon = 1e{-}9$. Further architectural and implementation details are provided in the Appendix.

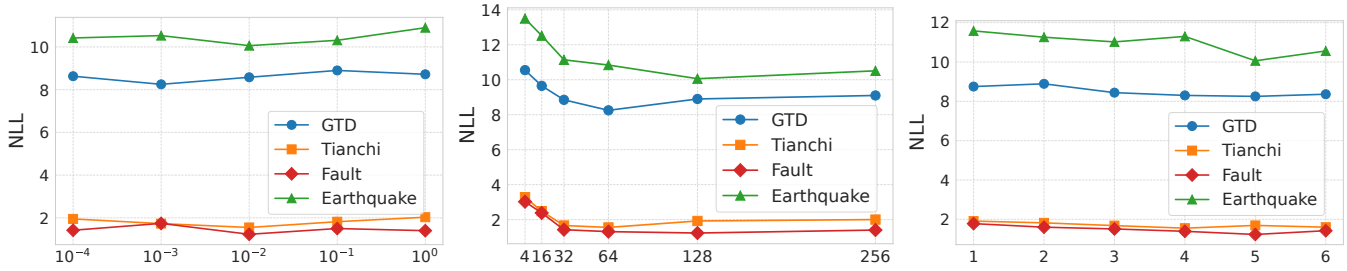
Experiments on Performance

Table 1 presents the performance of all models across four datasets using NLL, NMAE, and NRMSE as evaluation metrics. The proposed METP consistently achieves the best performance on the GTD dataset across all metrics. On the Tianchi dataset, it achieves the lowest NLL and NRMSE, while its NMAE is slightly higher than that of RMTPP, as the recurrent structure of RMTPP better captures short-term fluctuations in densely occurring events. On the Fault dataset, METP does not achieve the best NRMSE, which is slightly higher than that of RMTPP, due to RMSE's higher sensitivity to rare but large deviations, despite METP maintaining relatively strong overall accuracy across other metrics. For the Earthquake dataset, METP surpasses all baselines in NLL and NMAE, while its NRMSE remains marginally higher than that of ITHP, as ITHP demonstrates stronger suppression of long-interval prediction errors. Overall, METP obtains the lowest average scores across all three metrics, indicating superior robustness and generalization across diverse temporal event prediction tasks. Statistical tests further confirm that METP outperforms the baselines. Complete test results are reported in the Appendix.

Ablation Study

We perform ablation studies on both the GTD and Tianchi datasets to assess the contribution of each model component. As shown in Table 2, removing **external variables**

[3]https://www.kaggle.com/datasets/alessandrolobello/the-ultimate-earthquake-dataset-from-1990-2023/data



(a) Alignment regularization coefficient η (b) Scale of temporal encoding d (c) Number of granularities K

Figure 3: Performance variations of the method under different hyperparameter settings.

causes the most significant degradation, with NLL, NMAE, and NRMSE increasing by 15.6%, 28.9%, and 24.4% on the GTD dataset, and by 23.9%, 59.2%, and 21.3% on the Tianchi dataset, respectively. This highlights the critical role of environment-aware modeling. Excluding the **lagged weights** leads to an 8.6% rise in NLL on GTD and a 27.6% increase in NMAE on Tianchi, indicating the effectiveness of lagged temporal dynamics. When the **multi-scale structure** is ablated, performance deteriorates across all metrics, with a 7.3% increase in NRMSE on Tianchi, confirming its importance in capturing multi-granular temporal patterns. Finally, the **periodic structure** also contributes notably, especially on GTD where the NLL rises by 4.8%. These results demonstrate that each component plays a distinct role, and the full model consistently achieves the best performance across datasets.

Sensitivity Analysis

The sensitivity analysis evaluates METP's hyperparameters via NLL minimization. Fig. 3(a) shows optimal performance at $\eta = 10^{-3}$–10^{-2}, balancing feature alignment and flexibility since lower values under-regularize while higher values suppress informative variations. Fig. 3(b) indicates d = 64–128 optimally encodes temporal patterns, with smaller dimensions lacking expressiveness and larger ones causing overfitting. Fig. 3(c) reveals 4–5 granularity levels best capture multi-scale features, as fewer levels are insufficient and more introduce redundant computation. These results confirm METP's robustness when hyperparameters trade off representation capacity and model complexity.

Visualization of Event Prediction

Fig. 4 presents a comparative visualization of event prediction performance from January to June 2023. The two real-case examples demonstrate how historical external series evolve into future event time predictions, with METP effectively capturing the lagged temporal dependencies that influence event occurrences. Across the six predicted events (E1: 01-18 to E6: 05-28), METP's predictions show closer alignment with the actual events than those of THP and XTSFormer, especially during the prediction period where conventional methods tend to diverge. This performance advantage stems from METP's integrated modeling of historical event patterns, external variable influences, and multi-

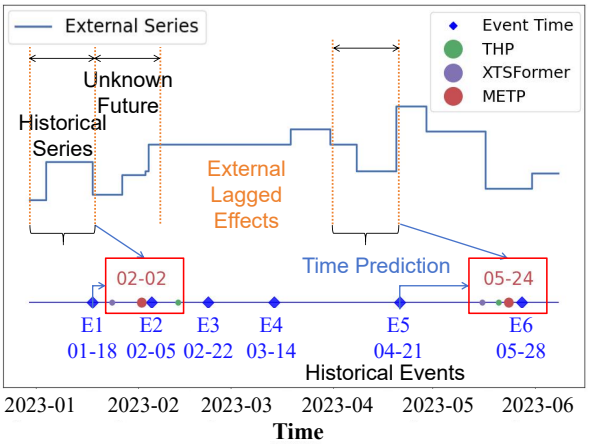


Figure 4: Comparison of event prediction between baselines and METP.

scale temporal relationships via its causal attention mechanism, enabling more stable and accurate forecasts throughout the prediction horizon. The visualization underscores that explicitly accounting for these external covariates allows METP to better anticipate the time of future events.

Conclusion and Future Work

The proposed METP framework advances temporal point process modeling by effectively incorporating multi-granular lagged external covariates through its novel lag-aware calibration and hierarchical fusion approach, demonstrating superior predictive accuracy across diverse datasets. Future extensions will focus on adapting the framework to streaming temporal point processes, enabling real-time event prediction under evolving external conditions.

Ethics Statement

The data used in this work, including public benchmarks and a proprietary dataset, consists of aggregated and de-identified information, ensuring no privacy violations. We acknowledge that enhanced forecasting models could be misused and strongly discourage such applications. Our

study focuses on general methodological improvement in temporal event modeling.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 72271009 and 72371217), the Guangzhou Industrial Informatics and Intelligence Key Laboratory (Grant No. 2024A03J0628), the Nansha Key Area Science and Technology Project (Grant No. 2023ZD003), and Project No. 2021JC02X191.

References

Boyd, A.; Bamler, R.; Mandt, S.; and Smyth, P. 2020. User-dependent neural sequence models for continuous-time event data. *Advances in Neural Information Processing Systems*, 33: 21488–21499.

Chen, S.; Xu, G.; Tao, T.; Zhang, S.; Zhang, K.; and Kuang, J. 2025. An Efficient Bearing Prognostic Approach through Modeling Multiperiodic and Nonperiodic Temporal Patterns. *IEEE Transactions on Industrial Informatics*.

Dai, Z.; He, L.; Yang, S.; and Leeke, M. 2024. Sarad: Spatial association-aware anomaly detection and diagnosis for multivariate time series. *Advances in Neural Information Processing Systems*, 37: 48371–48410.

Deng, J.; Ye, F.; Yin, D.; Song, X.; Tsang, I.; and Xiong, H. 2024. Parsimony or capability? decomposition delivers both in long-term time series forecasting. *Advances in Neural Information Processing Systems*, 37: 66687–66712.

Du, N.; Dai, H.; Trivedi, R.; Upadhyay, U.; Gomez-Rodriguez, M.; and Song, L. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1555–1564.

D'Amelio, A.; Cartella, G.; Cuculo, V.; Lucchi, M.; Cornia, M.; Cucchiara, R.; and Boccignone, G. 2025. TPP-Gaze: Modelling Gaze Dynamics in Space and Time with Neural Temporal Point Processes. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8786–8795. IEEE.

Gao, J.; Cao, Q.; and Chen, Y. 2025. Auto-regressive moving diffusion models for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16727–16735.

Gracious, T.; and Dukkipati, A. 2025. Deep Representation Learning for Forecasting Recursive and Multi-Relational Events in Temporal Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16897–16905.

Gu, Y. 2021. Attentive neural point processes for event forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7592–7600.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1): 83–90.

Isik, Y.; Chapfuwa, P.; Davis, C.; and Henao, R. 2023. Hawkes Process with Flexible Triggering Kernels. In *Machine Learning for Healthcare Conference*, 308–320. PMLR.

Karpukhin, I.; Shipilov, F.; and Savchenko, A. 2024. HoTPP Benchmark: Are We Good at the Long Horizon Events Forecasting? *arXiv preprint arXiv:2406.14341*.

Kuang, Y.; Yang, C.; Yang, Y.; and Li, S. 2024. Unveiling latent causal rules: A temporal point process approach for abnormal event explanation. In *International Conference on Artificial Intelligence and Statistics*, 2935–2943. PMLR.

Li, Y.; Li, K.; Chen, C.; Zhou, X.; Zeng, Z.; and Li, K. 2021a. Modeling temporal patterns with dilated convolutions for time-series forecasting. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16: 1–22.

Li, Z.; Zhao, Y.; Han, J.; Su, Y.; Jiao, R.; Wen, X.; and Pei, D. 2021b. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 3220–3230.

Lin, S.; Lin, W.; Hu, X.; Wu, W.; Mo, R.; and Zhong, H. 2024. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37: 106315–106345.

Liu, P.; Guo, H.; Dai, T.; Li, N.; Bao, J.; Ren, X.; Jiang, Y.; and Xia, S.-T. 2025. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18915–18923.

Meng, Z.; Wan, K.; Huang, Y.; Li, Z.; Wang, Y.; and Zhou, F. 2024. Interpretable Transformer Hawkes processes: Unveiling complex interactions in social networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2200–2211.

Mu, Y.; Shahzad, M.; and Zhu, X. X. 2025. MPTSNet: Integrating multiscale periodic local patterns and global dependencies for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19572–19580.

Shams Eddin, M. H.; and Gall, J. 2024. Identifying spatio-temporal drivers of extreme events. *Advances in Neural Information Processing Systems*, 37: 93714–93766.

Shchur, O.; Biloš, M.; and Günnemann, S. 2019. Intensity-free learning of temporal point processes. *arXiv preprint arXiv:1909.12127*.

Shou, X.; Gao, T.; Subramanian, D.; Bhattacharjya, D.; and Bennett, K. P. 2023. Concurrent multi-label prediction in event streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9820–9828.

Xiao, T.; Xu, Z.; He, W.; Xiao, Z.; Zhang, Y.; Liu, Z.; Chen, S.; Thai, M. T.; Bian, J.; Rashidi, P.; et al. 2025. XTS-Former: Cross-Temporal-Scale Transformer for Irregular-Time Event Prediction in Clinical Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28502–28510.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*.

Yang, C.; Mei, H.; and Eisner, J. 2022. Transformer Embeddings of Irregularly Spaced Events and Their Participants. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*.

Zhang, D.; Du, C.; Peng, Y.; Liu, J.; Mohammed, S.; and Calvi, A. 2024a. A multi-source dynamic temporal point process model for train delay prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Zhang, P.; Zhou, Z.; Feng, Z.; Wang, J.; and Zhang, Y. 2023. Inference and analysis on the evidential reasoning rule with time-lagged dependencies. *Engineering Applications of Artificial Intelligence*, 126: 106978.

Zhang, Q.; Lipani, A.; Kirnap, O.; and Yilmaz, E. 2020. Self-attentive Hawkes process. In *International conference on machine learning*, 11183–11193. PMLR.

Zhang, S.; Zhou, C.; Liu, Y. A.; Zhang, P.; Lin, X.; and Ma, Z.-M. 2024b. Neural jump-diffusion temporal point processes. In *Forty-first International Conference on Machine Learning*.

Zhang, W.; Zhang, L.; Han, J.; Liu, H.; Fu, Y.; Zhou, J.; Mei, Y.; and Xiong, H. 2024c. Irregular traffic time series forecasting based on asynchronous spatio-temporal graph convolutional networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4302–4313.

Zuo, S.; Jiang, H.; Li, Z.; Zhao, T.; and Zha, H. 2020. Transformer hawkes process. In *International conference on machine learning*, 11692–11702. PMLR.