A DFT-Based Periodicity Extraction

In the **Periodicity and Multi-Granular Encoding** subsection of Section **Methodology**, we extract dominant periodic patterns from the external covariate time series $\{x_1, x_2, \ldots, x_t\}$ by applying the Discrete Fourier Transform (DFT), which converts the input from the time domain into the frequency domain:

$$G(f) = \sum_{n=1}^{t} x_n e^{-j2\pi fn/t}, \tag{A.1}$$

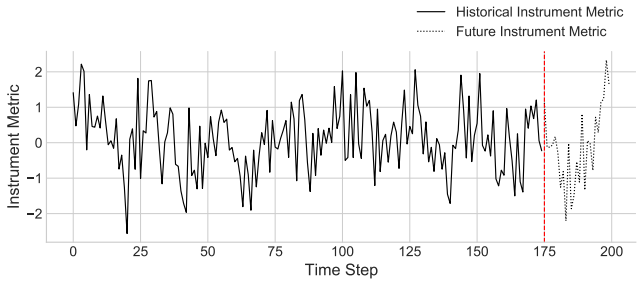
where f denotes the frequency index and $j = \sqrt{-1}$ is the imaginary unit. The spectral magnitude $|G(f)|$ quantifies the strength of each frequency component. To identify the dominant cyclic behavior, we exclude the DC (zero-frequency) component and select the frequency with the largest amplitude:

$$f^* = \arg\max_{f \neq 0} |G(f)|, \quad \delta = \frac{1}{f^*}, \tag{A.2}$$

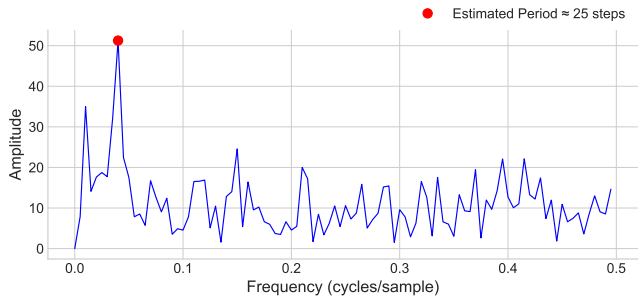
where δ denotes the estimated dominant period in the original time domain. A sliding window of length δ is then applied to extract the most recent segment:

$$x_{t-\delta+1:t} = [x_{t-\delta+1}, \ldots, x_t]. \tag{A.3}$$

To illustrate this procedure, we take the external covariate representing *motor current*, which is closely related to fault occurrences, as an example. This variable is used to reveal periodic patterns that may indicate underlying mechanical degradation. The analysis is visualized in Fig. 1. Figure 1(a) illustrates the time series of *motor current* collected from sensors, partitioned at timestamp T into a historical (retrospective) window and an unknown future (prospective) window. The historical segment serves as the basis for subsequent periodicity extraction. Figure 1(b) shows the corresponding frequency spectrum obtained via DFT. The spectral peak indicates the dominant periodicity, which is then used to inform model input. The extracted frequency-aware representation is incorporated into the predictive model to enhance its ability to capture temporal regularities associated with fault patterns.



(a) Time series of the external covariate motor current, collected from sensor instruments, based on historical data at the timestamp.



(b) DFT spectrum of the historical segment, highlighting the dominant frequency component.

Figure 1: DFT-based periodicity analysis of the external covariate (motor current): (a) raw time series at timestamp T; (b) spectral representation used to estimate the dominant period.

B Algorithm

As described in the **Training Objective** subsection of Section **Methodology**, a *two-stage optimization strategy* is employed to improve training stability by explicitly decoupling temporal structure learning from intensity modeling.

Stage I: Temporal Structure Training (Inner Loop). In the first stage, an inner optimization loop is employed to focus on minimizing losses $\mathcal{L}_k$, which enforce alignment between the attention weights and predefined reversed geometric priors. This encourages the attention mechanism to identify and encode meaningful temporal dependencies within the external covariates, independently of the prediction task. During this stage, all parameters associated with the event intensity function are held fixed, effectively isolating the learning of temporal structures from event occurrence modeling.

Stage II: Joint Optimization of Full Model. In the second stage, the full set of model parameters is jointly optimized by minimizing the total loss $\mathcal{L}$, which integrates both the event prediction loss and the auxiliary alignment losses. The attention mechanism, regularized during Stage I, operates as a soft selector, incorporating the learned temporal patterns into the event intensity modeling process. This facilitates the learning of coherent temporal representations while preserving interpretability, ultimately enhancing the accuracy of event forecasting and improving anomaly localization performance. The complete training procedure is outlined in Algorithm 1.

Algorithm 1: Two-Stage Training Procedure of METP

Require: Historical event sequence $\mathcal{S}_{1:n} = \{t_1, \ldots, t_n\}$; External covariates $\mathcal{X}_T = \{x_1, \ldots, x_T\}$; Granularity set $\{s_k\}_{k=1}^{K}$; trade-off coefficient η;

Ensure: Intensity $\hat{\lambda}(t)$ and predicted next event time $\hat{t}_{n+1}$; Inner-loop loss $\{L_k = 0\}_{k=1}^{K}$, Log-likelihood loss L_p=0;

Identify dominant period δ from $\mathcal{X}_T$ via Eq. (A.2);
for every time $t \leq T$ **do**
  Recent covariate segment $x_{t-\delta+1:t} = [x_{t-\delta+1}, \ldots, x_t]$ and periodic matrix $\mathbf{X}(t) = [x_{t-\delta+1:t}, \ldots,]^\top$ (Eq. 5);
  for each granularity s_k **do**
    Pooled feature: $\mathbf{U}_k(t) = \mathrm{MaxPool}_{s_k}(\mathbf{X}(t))$ (Eq. 6);
  end for
end for
for event $j = 1$ to n **do**
  Sinusoidal $e(t_j)$ for event time t_j (Eq. 7);
end for
Stage I: Temporal Structure Learning
for granularity $k = 1$ to K **do**
  while $\mathcal{L}_k$ not converged **do**
    Query, key vectors and decay rate: $q_k = \mathbf{u}_k\mathbf{w}_1$, $k_k = \mathbf{u}_k\mathbf{w}_2$ and $\phi_k = \mathbf{u}_k\mathbf{w}_k$;
    for time $j = 1, \ldots, \gamma_k$ **do**
      Reverse geometric prior (Eq. 8): $r_k(j) = \frac{(1-\phi_k)^{j-1}\phi_k}{1-(1-\phi_k)^{\gamma_k}}$;
      Raw attention score $s_k^j = \frac{q_k^\top k_k^j}{\sqrt{l}}$ and normalized attention score and distribution (Eqs. 9–10):

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

    end for
    Symmetric KL divergence: $\mathrm{D}_{\mathrm{SKL}}(k) = \mathrm{D}_{\mathrm{KL}}(p_k \,\|\, r_k) + \mathrm{D}_{\mathrm{KL}}(r_k \,\|\, p_k)$ (Eqs. 11–12);
    Weighted features: $\mathbf{z}_t^k = \sum_{j=1}^{\gamma_k}\alpha_k^j \mathbf{u}_k^{t-j+1}$ (Eq. (13));
    Event label $\hat{y}_t^k = \sigma(w_z\mathbf{z}_t^k + b_z)$ and Inner-loop loss (Eq. 14-15):

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

  end while
  for every time $t \leq T$ **do**
    Aligned external representation: $\mathbf{P}_k(t) = \frac{1}{Z_k}\sum_i \alpha_k^i \cdot d_k(t-t_j)\cdot \mathbf{u}_k^{t_j-i+1}$ (Eq. 16);
  end for
end for
Stage II: Joint Optimization
while $\mathcal{L}$ not converged **do**
  Hierarchical MoE $\mathbf{P}(t) = \pi_1\mathbf{P}_1(t) + \sum_{k\in\mathcal{K}_t}\pi_k\mathbf{P}_k(t)$ and event embeddings $\mathbf{H}$ (Eqs. 17–19);
  Conditional intensity (Eq. 20):

$$\lambda(t) = \mathrm{softplus}\left(\rho\cdot\frac{t-t_i}{t_i} + \mathbf{w}_{\mathrm{a}}^\top[\mathbf{h}_t, \mathbf{p}(t)] + b\right)$$

  Log-likelihood loss (Eq. 22):

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

  Total loss: $\mathcal{L} = \mathcal{L}_p + \sum_{k=1}^{K}\mathcal{L}_k$ (Eq. 23);
end while
Prediction: Optimized $\hat{\lambda}(t)$ and next event time estimated by: $\hat{t}_{n+1} = \sum_{t=T}^{M} t\cdot\kappa(t)$ (Eq. 24).

C Datasets

To ensure robust evaluation across diverse application domains, both proprietary and public datasets are employed in the **Experiment Settings** subsection of Section **Experiments**.

Fuel Transaction Dataset (Proprietary). This dataset comprises 3,592 fuel transaction records collected from gasoline stations between August 2022 and April 2024. Each record contains a customer identifier, transaction timestamp, and expenditure amount. For temporal point process analysis, we focus on customers with more than three recorded transactions. The external covariate used is the transaction price at each time point, which influences consumer purchasing behavior by modulating the timing and frequency of transactions. In accordance with binding confidentiality agreements and relevant data protection laws, public release of the full proprietary dataset is not permitted. We provide comprehensive statistical summaries in Table 1 to support the reproducibility and interpretability of our results within the permissible scope.

Tianchi-Walmart Storm Sales Dataset[1] contains daily sales records for 111 products across 45 retail stores, spanning 1,843 storm events. This dataset is enriched with weather information from NOAA, providing comprehensive environmental context. Given the strong weather sensitivity of the included products, we select temperature as the external variable to capture its effect on purchasing dynamics. Temperature fluctuations are known to significantly impact consumer demand patterns during and following extreme weather conditions, thus serving as a critical temporal covariate in modeling storm-related sales behavior.

Elevator Fault Dataset[2] comprises high-frequency sensor data sampled at 4 Hz over a 30-day monitoring period from 10 elevators. Fault occurrences are explicitly annotated as events. To effectively capture temporal dynamics, the raw data are segmented into minute-aligned sequences. Selected sensor readings are used as external covariates, as they reflect real-time operational states of the elevators. These metrics facilitate early fault detection by revealing subtle deviations in system behavior, thus supporting predictive maintenance and anomaly localization.

Global Earthquake Dataset[3] contains approximately 3 million seismic records globally from 1990 to 2023. Each entry includes detailed attributes such as magnitude, depth, geographical coordinates, and intensity metrics, making it well-suited for large-scale geophysical risk modeling. In addition, we incorporate corresponding regional weather conditions as external covariates, as environmental factors may influence both the manifestation of seismic events and their secondary effects. This contextual integration enhances the model's predictive capability and interpretability in spatiotemporal earthquake forecasting.

Dataset	Records	Time Span	Event Type	Covariate
Fuel Transaction	36, 575	Aug 2022 – Apr 2024	Fuel Purchase	Transaction Price
Tianchi-Walmart Storm	118, 696	Jan 2016 – Dec 2018	Product Sale	Temperature
Elevator Fault	164, 517	30 Days	Fault Alert	Sensor Readings
Global Earthquakes	3361, 846	Jan 1990 – Dec 2023	Seismic Event	Local Weather

Table 1: Summary statistics of datasets used in experiments, including event types and associated covariates.

D Implementation Details

In the **Implementation Details** subsection of the **Experiments** section, we present the technical specifications and ablation settings of the proposed method. The entire model is implemented in PyTorch using a modular design tailored for next-event time prediction under a single-event-type scenario. All experiments are conducted on a workstation equipped with an Intel Core i5 CPU and an NVIDIA RTX 3070 GPU, which provides adequate computational resources for both training and evaluation. The implementation is based on Python and utilizes several key libraries, including `torch==1.10.2`, `numpy==1.19.3`, `matplotlib==3.3.4`, and `scikit_learn==1.7.1`, which support model construction, training, visualization, and evaluation.

The model is trained using the Adam optimizer, with an initial learning rate set to 1×10^{-3}. The choice of Adam facilitates efficient gradient-based optimization by adaptively adjusting the learning rates of individual parameters, thereby accelerating convergence and improving stability. Training is performed with a batch size of 16, which balances the trade-off between computational efficiency and gradient estimation accuracy. The entire training process runs for 100 epochs, a duration empirically determined to ensure convergence without overfitting on the training data. To maintain numerical stability during the calculation of the Kullback–Leibler (KL) divergence term in the loss function, a small constant $\epsilon = 10^{-9}$ is added. This prevents potential issues arising from logarithms of zero or extremely small probabilities, thus avoiding gradient explosion or vanishing problems during backpropagation. The dominant temporal period δ of input sequences is estimated using a Fast Fourier Transform (FFT)-based periodogram analysis. This method effectively captures periodic patterns by identifying the frequency component with the highest power spectrum density, which is crucial for modeling temporal dependencies in event sequences. Unless otherwise

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

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

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

specified, the embedding dimension for all input features and intermediate representations is fixed at 64, providing a rich yet computationally manageable latent space. The temporal encoder incorporates $K = 5$ parallel max-pooling branches, each operating with distinct stride lengths $s \in \{1, 2, 4, 6, 8\}$. This hierarchical pooling strategy enables the model to extract multi-scale periodic features, capturing both fine-grained and coarse temporal patterns. Additionally, the model maintains a memory depth of $m = 3$, which controls the temporal context length used for prediction, balancing the complexity and responsiveness of the learned temporal dependencies. The code examples and corresponding guidance documents are included in the supply file, provided separately for clarity and reproducibility.

E Statistical Significance Testing

In the **Experiments on Performance** subsection of Section **Experiments**, we validate the significance of our method over the baselines with statistical tests, employing both parametric (paired t-test) and non-parametric (Friedman–Nemenyi test) methods across the four datasets.

E.1 Paired *t*-Test

We conduct a paired t-test on the performance metrics (e.g., RMSE, accuracy) collected from n repeated runs. Let x_i and y_i denote the evaluation scores of our method and the benchmark in the i-th trial. Define the difference as $d_i = x_i - y_i$, then compute the test statistic as:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}, \quad t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{A.4}$$

where $\bar{d}$ and s_d are the sample mean and standard deviation of the differences. The resulting t-statistic is evaluated under a t-distribution with $n - 1$ degrees of freedom. The paired t-test is a statistical method used to evaluate whether the performance improvement of one model over another is statistically significant, by comparing their results under repeated experimental trials. A larger absolute value of the t-statistic generally indicates stronger evidence against the null hypothesis of no performance difference. In this work, we conducted a paired t-test to analyze the performance differences between the METP model and the average results of seven baseline models. The test was performed across key evaluation metrics, including RMSE (where lower values indicate better prediction accuracy) and accuracy (where higher values are preferred). The results demonstrate that METP significantly outperforms the baselines, indicating its superior predictive capability and robustness across evaluated metrics. The results demonstrate that METP significantly outperforms the baselines, indicating its superior predictive capability and robustness across evaluated metrics, as shown in Figure 2.

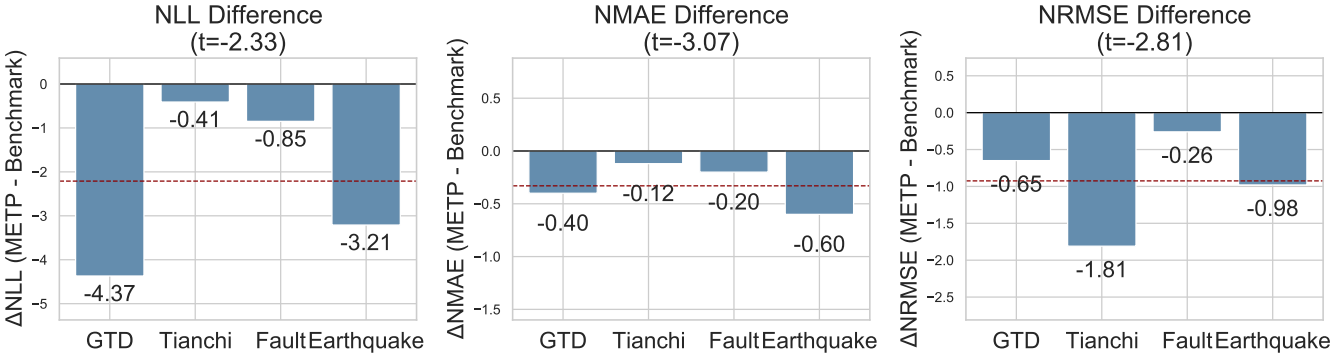


Figure 2: Paired t-test results comparing the METP model against the average of seven benchmark models across four datasets. Each subfigure shows the performance difference Δ for a specific metric (lower is better). The dashed red line indicates the mean difference across datasets. Negative bars suggest that METP outperforms the baseline.

E.2 Friedman and Nemenyi Tests

To compare multiple models across N datasets or experimental settings, we adopt the Friedman test. Let $R_{i,j}$ be the rank of the j-th method on the i-th dataset. The Friedman statistic is given by:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k}\left(\bar{R}_j - \frac{k+1}{2}\right)^2\right], \tag{A.5}$$

where $\bar{R}_j = \frac{1}{N}\sum_{i=1}^{N} R_{i,j}$. If χ_F^2 is significant, we conduct the Nemenyi post-hoc test. The critical difference (CD) between two average ranks is calculated as:

$$\mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{A.6}$$

where q_α is the critical value from the Studentized range distribution. A difference in average ranks larger than CD indicates statistical significance. This indicator reflects the overall difference in model rankings across datasets. Smaller average ranks correspond to better model performance, so the goal is to have lower rank values. The Friedman test yields a statistically significant result ($p < 0.05$), suggesting that the model performs differently from the others. According to the Nemenyi post-hoc test (Figure 3), METP demonstrates consistently superior performance over all baseline models across the evaluated datasets.

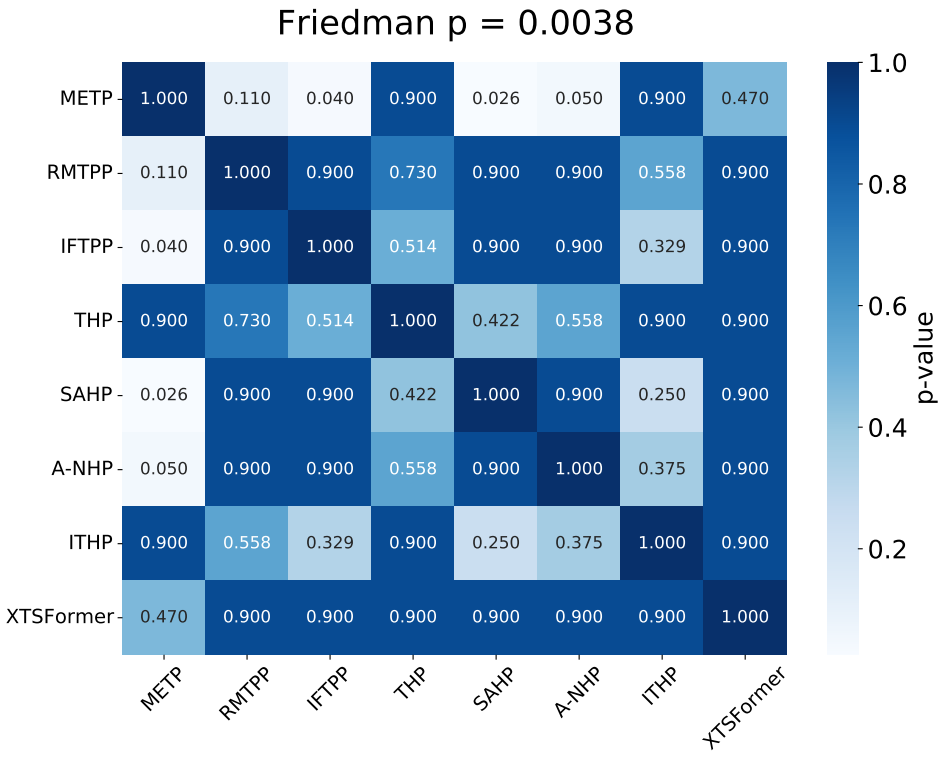


Figure 3: Nemenyi post-hoc test results based on the Friedman test across four datasets. Darker cells indicate more significant differences (lower p-values) in pairwise model comparisons. The diagonal values are all 1, indicating perfect self-comparison for each model.