I. Derivation of TWFA as a Generalized Form of TEO

First, we begin with the original definition of the attention score at i-th head in t-th time point:

$$
\begin{aligned}
M_{i,t} &= (K_{i,t} - V_{i,t})^2 \\
&= (X_{i,t+1}W_{i,t}^K - X_{i,t-1}W_{i,t}^V)^2 \\
&= (X_{i,t+1}W_{i,t}^K)^2 - 2(X_{i,t+1}W_{i,t}^K)(X_{i,t-1}W_{i,t}^V) \\
&\quad + (X_{i,t-1}W_{i,t}^V)^2.
\end{aligned} \tag{1}
$$

To simplify the derivation, we assume that the weight matrices $W_{i,t}^K$ and $W_{i,t}^V$ are equal, denoted as $W_{i,t}$:

$$W_{i,t}^K = W_{i,t}^V = W_{i,t}. \tag{2}$$

This assumption is reasonable when the transformations of the key and value are similar or shared. Substituting into Eq. (1), we obtain:

$$
\begin{aligned}
M_{i,t} &= (X_{i,t+1}W_{i,t})^2 - 2(X_{i,t+1}W_{i,t})(X_{i,t-1}W_{i,t}) \qquad (3)\\
&\quad + (X_{i,t-1}W_{i,t})^2 \\
&= W_{i,t}\left(X_{i,t+1}^2 - 2X_{i,t+1}X_{i,t-1} + X_{i,t-1}^2\right)W_{i,t}. \qquad (4)
\end{aligned}
$$

The expression in parentheses in Eq. (4) is a standard expansion:

$$X_{i,t+1}^2 - 2X_{i,t+1}X_{i,t-1} + X_{i,t-1}^2 = (X_{i,t+1} - X_{i,t-1})^2. \tag{5}$$

For the discrete signal X, we can utilize forward and backward differences. Assuming the signal X changes smoothly at position i, t, we approximate $X_{i,t+1}$ and $X_{i,t-1}$ as follows:

$$X_{i,t+1} \approx X_{i,t} + \Delta X_{i,t}, \tag{6}$$

$$X_{i,t-1} \approx X_{i,t} - \Delta X_{i,t}, \tag{7}$$

where $\Delta X_{i,t}$ represents the difference at position i, t, and for a discrete case, the step size can be assumed to be 1.

Calculating the Squares of $X_{i,t+1}$ and $X_{i,t-1}$

Using Eq. (6) and Eq. (7), we calculate the squares of $X_{i,t+1}$ and $X_{i,t-1}$:

1. Squares of $X_{i,t+1}$ and $X_{i,t-1}$:

$$
\begin{aligned}
X_{i,t+1}^2 &\approx (X_{i,t} + \Delta X_{i,t})^2 && (8)\\
&= X_{i,t}^2 + 2X_{i,t}\Delta X_{i,t} + (\Delta X_{i,t})^2, && (9)\\
X_{i,t-1}^2 &\approx (X_{i,t} - \Delta X_{i,t})^2 && (10)\\
&= X_{i,t}^2 - 2X_{i,t}\Delta X_{i,t} + (\Delta X_{i,t})^2. && (11)
\end{aligned}
$$

2. Calculating the product $X_{i,t+1}X_{i,t-1}$:

$$
\begin{aligned}
X_{i,t+1}X_{i,t-1} &\approx (X_{i,t} + \Delta X_{i,t})(X_{i,t} - \Delta X_{i,t}) \\
&= X_{i,t}^2 - (\Delta X_{i,t})^2.
\end{aligned} \tag{12}
$$

Substituting and Simplifying

Substituting Eq. (9), Eq. (11), and (12) into Eq. (5), we get:

$$
\begin{aligned}
M_{i,t} &\approx W_{i,t}\Big[\big(X_{i,t}^2 + 2X_{i,t}\Delta X_{i,t} + (\Delta X_{i,t})^2\big) \\
&\quad - 2\big(X_{i,t}^2 - (\Delta X_{i,t})^2\big) \\
&\quad + \big(X_{i,t}^2 - 2X_{i,t}\Delta X_{i,t} + (\Delta X_{i,t})^2\big)\Big]W_{i,t} \\
&= W_{i,t}\big[2(\Delta X_{i,t})^2\big]W_{i,t}.
\end{aligned} \tag{13}
$$

According to Eq. (12), we can express $(\Delta X_{i,t})^2$ in terms of $X_{i,t}^2$ and $X_{i,t+1}X_{i,t-1}$:

$$(\Delta X_{i,t})^2 \approx X_{i,t}^2 - X_{i,t+1}X_{i,t-1}. \tag{14}$$

Substituting Eq. (14) into Eq. (13), we obtain:

$$T_{i,t} \approx 2W_{i,t}\left(X_{i,t}^2 - X_{i,t+1}X_{i,t-1}\right)W_{i,t}. \tag{15}$$

The final Eq. (15) represents a generalized form of the TEO mechanism, where the attention score $M_{i,t}$ is formulated in terms of the squared values of the signal and the interaction between neighboring points.

II. Detailed Description of Baseline Models

1) Crossformer [30]: Divides fault data into patches using Dimension-Segment-Wise Embedding and applies attention in both time and feature dimensions through a Two-Stage Attention Layer. Fault features are extracted via a Hierarchical Encoder-Decoder mechanism for detection.
2) ETSformer [37]: Uses temporal convolution filters and multi-head exponential smoothing attention to extract fault features. Growth and seasonality are modeled through stacked modules, and the decoder further processes these features for fault detection.
3) FEDformer [38]: Uses an encoder-decoder structure, with a Period-Trend Decomposition module to split sequences into trend and periodic components. The encoder focuses on periodic components using frequency attention, while the decoder extracts fault features.
4) Informer [31]: Uses a Transformer-based Encoder, embedding fault data and applying ProbSparse multi-head self-attention and attention distillation in multiple layers to output fault features.
5) MICN [32]: Decomposes sequences into seasonal and trend-period components using a multi-scale decomposition module. These components are modeled with MIC Multi-scale Isometric Convolution layers and linear regression to generate fault features.
6) Pyraformer [33]: Constructs multi-resolution trees using Coarse-Scale Construction Module . Fault features are generated through pyramid attention, residual modules, and feedforward networks in multiple encoder layers.
7) Transformer [34]: Uses multiple Encoder layers with multi-head self-attention, residual connections, and feedforward layers to extract features. Masked and cross-attention mechanisms are applied in the decoder for final fault features.
8) TimesNet [35]: Uses Fourier Transform to convert time series to the frequency domain. Extracts key frequency

information and builds 2D time series, using parameter-efficient inception blocks and residual connections to extract fault features.

9) DCNN-Transformer [36]: Applies 1D deep CNN to extract features, followed by a Transformer encoder for sequence learning. Uses attention mechanisms to capture long-term dependencies for fault detection.
10) TP-FCN [27]: Encodes and segments fault signals, creating ZSV-ZSC images. Uses a dual-path fully convolutional network to estimate fault initiation time and extent.
11) DC-CNN [39]: Uses STFT to extract frequency features from fault signals. Constructs time-frequency feature images and inputs them to a fully connected layer with maxout units for fault detection.