JAISCR, 2026, Vol. 16, No. 4, pp. 397 – 419
10.2478/jaiscr-2026-0019

# ANOMALY PREDICTION METHOD FOR COMPLEX SCENARIOS BASED ON MULTI-MODAL CAUSAL INTERVENTION

Ruoyuan Zhang[1,2,3], Xianwen Fang[1,2,3,*], Ku Lu[1,2,3]

[1]*School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, 232000, China*

[2]*Anhui Water Conservancy Technical College, Hefei, 230000, China*

[3]*Anhui Province Engineering Laboratory for Big Data Analysis and Early Warning Technology of Coal Mine Safety, Huainan, 232000, China*

[*]*E-mail: xwfang@aust.edu.cn*

*Submitted: 28th February 2026; Accepted: 19th May 2026*

**Abstract**

Anomaly prediction is vital for ensuring the safe operation of pump station facilities. However, traditional models are often hampered by spurious correlation stemming from unobservable confounding factors, thereby degrading detection accuracy. To address this, this paper proposes the Causal Intervention with Front-door Adjustment Model (CIFAM) to eliminate spurious correlation in complex features. Initially, CIFAM integrates video and signal data, utilizing graph attention networks and clustering to construct a multi-level structural attribute graph that represents dynamic operational dependencies. The model incorporates causal dilated convolution to capture long-term memory while preventing information leakage, alongside a random edge dropping strategy to simulate causal intervention. By applying the front-door adjustment criterion, CIFAM extracts robust intermediate representations to block interference paths. Furthermore, contrastive learning and Adaptive Instance Normalization (AdaIN) are employed to decouple features, bolstering model robustness in complex scenarios. Experimental results demonstrate that CIFAM achieves state-of-the-art performance across multiple industrial benchmarks, specifically reaching 91.1% accuracy and a 90.6% Micro-F1 score on the OURS dataset. The model exhibits exceptional robustness and sample utilization efficiency. Furthermore, with minor modifications, this algorithm can be generalized to various monitoring scenarios, such as power generator set monitoring, semiconductor production lines, and UAV-assisted monitoring for oil pipelines or high-voltage power line maintenance.

**Keywords:** pump station anomaly prediction, multimodal fusion, front door adjustment strategy, causal intervention, causal dilated convolution

# 1 Introduction

The key task to ensure the safe and stable operation of systems in complex scenarios such as pump stations and large machine rooms is to do a good job in industrial anomaly prediction [1, 2]. With the acceleration of the Industrial 4.0 process, the equipment structure becomes more precise and the mechanism more complex. Once any minor initial anomaly occurs, it is prone to rapid spread due to physical coupling and trigger other secondary sudden failures, which can easily lead to safety accidents [3]. Therefore, how to use multi-sensor technology to achieve accurate fault early warning has become a key research object in recent years to ensure the safe and stable operation of industrial control systems. Against this background, most of the anomaly detection methods proposed in the early stage are based on the physical quantity detection of single-modal sensors. However, in the actual industrial production environment, it is difficult for a single modality to achieve high coverage [4], and most industrial process anomaly scenarios mainly obtain real-time data through monitoring videos [5]. Therefore, how to combine visual, sensor information and other channels to achieve multi-modal early warning of industrial faults has become a hot topic in current industrial early warning research. Due to the extensive use of non-invasive monitoring methods in actual industrial scenarios, there are problems of "semantic gap" between multi-modal heterogeneous signals and very complex nonlinear interactions and dynamic topologies in different operation activities [6]. Although Spatio-Temporal Graph Neural Networks (STGNN) have a good effect in capturing spatio-temporal dependencies [7], there are still some fundamental problems in actual multi-modal prediction scenarios, that is, the model often focuses on statistical correlations rather than causal logic in the data generation process. In the actual industrial environment, there are many unobservable confounding factors [8], such as sudden changes in environmental lighting, random drift of sensors over time, or potential arbitrary operation methods by humans, which can lead to both the observed data and the prediction target being affected and introducing spurious correlation. If only empirical risk minimization learning is adopted, the model is prone to only capture the shortcut features of the problem rather than finding the cause of such results [9, 10]. That is, if it is found in the training set samples that changes in lighting and equipment shutdowns occur simultaneously, it may be concluded that there is a correlation between the strength of lighting and whether the equipment is abnormal. In this case, the prediction accuracy will be affected when facing new working conditions and new sites, that is, it cannot robustly predict the data in the industrial site. To solve these problems, causal reasoning provides an effective theoretical guidance for the model to transform from associative logic to causal logic [11]. Under the premise of observing a small number of confounding factors, traditional backdoor adjustment rules are difficult to adjust all factors [12]; in contrast, the method based on frontdoor adjustment has higher superiority by introducing intermediate variables to block the confounding paths. Therefore, we propose the CIFAM model, which introduces a contrastive learning mechanism to perform frontdoor causal intervention. In response to the characteristics of asynchronous anomaly signals in pump station working conditions, the Causal Dilated Convolution module is introduced to obtain more historical state information from earlier time steps without increasing computational complexity, thereby ensuring that the model can truly learn the interaction logic between nodes from the real causal sequence and eliminate the spurious correlation caused by unobserved interference factors from the source. By forcing the extraction of invariant intermediate representations under local disturbances in the dual-view interaction, the spurious correlation caused by unobservable confounding factors is stripped away. At the same time, the AdaIN technology is combined with the self-training fusion prior initialization method, and the distribution shift caused by the change in working conditions is smoothed through adaptive instance normalization, ensuring that the model's anomaly prediction always follows the true logical evolution path. The main contributions of this paper are as follows:

1. To address the issue of spurious correlations, a multimodal causal intervention framework based on front-door adjustment is proposed. Unlike the back-door adjustment that relies on observable variables, this paper utilizes mediator variables to identify the true causal effect, theoretically resolving the prediction bias caused by

unobservable confounding factors.

2. To tackle signal delay and information leakage risks, an expanded convolution module with causal constraints is designed. By exponentially expanding the receptive field, the model can effectively trace early operational parameters and enhance long-term memory capacity while strictly adhering to the physical time sequence.

3. To deal with feature interference and distribution shift, contrastive learning and AdaIN techniques are combined. By randomly dropping edges to simulate the intervention process, the model is forced to extract invariant representations robust to local perturbations, achieving a deep decoupling of core causal content from environmental noise.

4. Adaptive instance normalization technology is utilized to achieve deep decoupling of features. Verified on a real pumping station platform and multiple public datasets, the proposed method demonstrates excellent performance in suppressing spurious correlation and improving prediction accuracy.

# 2 Related Work

To provide a comprehensive background for the proposed CIFAM, this section reviews the existing literature from three key dimensions: industrial anomaly prediction, the development of graph neural networks, and causal intervention and reasoning.

## 2.1 Research on anomaly prediction

Sensor signals in industrial scenarios are typically multimodal, such as vibration signals, current signals, and other microscopic one-dimensional signals related to the operation of pump stations, as well as macroscopic two-dimensional image and video signals representing operational and environmental conditions [13]. These signals have different semantics. Deep learning can effectively extract global features from the data, and experiments have shown that it can conveniently classify these heterogeneous data. The initial methods used convolutional neural networks (CNNs) to extract local periodicity from monitoring signals and spatial semantics from video frames. For time series data, these methods often generated time-frequency images to capture underlying patterns. Meanwhile, for video streams, they directly extracted visual features to perform abnormal activity analysis and recognition [14]. For example, Liu et al. [15] proposed a multivariate convolutional neural network that can effectively extract the correlation of lagged information described in sensor signals. For non-intrusive monitoring based on video, it uses CNN-based object detection to identify certain behaviors or abnormal operations. Recurrent neural networks and some of their variants (such as LSTM and GRU) are designed to model temporal dependencies, and recurrent neural network models are used to perform sequence regression on single-modal equipment health status and predict sequences based on the length of the time domain. Additionally, some studies combine local Fisher discriminant methods with other statistical methods to fuse and reduce the dimensionality of these high-dimensional heterogeneous data, improving the effectiveness and efficiency of classifiers [16]. Overall, all the above methods extract activity features at individual time steps separately, but they usually treat the signals at individual time steps as independent features or simply stack multiple time steps' signals linearly [17]. This approach essentially ignores the strict process nature of pump station operations, where different operational activities have close logical connections rather than simple linear interactions. However, existing fusion methods mostly use linear stacking or simple concatenation of features, neglecting the strict process logic between operational activities during pump station operations. This processing method cannot effectively capture cross-stage activity-level interactions, affecting the reliability of the model in predicting complex process anomalies.

## 2.2 Graph neural network modeling

When dealing with spatio-temporal data in complex scenarios such as industrial pumping stations or transportation networks, effectively capturing the complex nonlinear interactions and dynamic topological associations among variables is crucial for achieving accurate predictions. Early researchers introduced Recurrent Neural Networks (RNN) and their variants, such as Long Short-

Term Memory networks (LSTM) and Gated Recurrent Units (GRU), to capture long-range temporal dependencies in multi-dimensional time series [18]. Meanwhile, some scholars adopted Convolutional Neural Networks (CNN) to identify spatial latent correlations in grid-based systems [19]. However, to address the complex coupling relationships in non-Euclidean spaces, many researchers constructed Graph Neural Networks (GNN) to effectively extract the complex spatial correlations among device components or sensor nodes [20]. In the field of graph convolutional modeling, Kipf et al. [21] proposed a first-order approximation scheme of Graph Convolutional Networks (GCN) to achieve efficient diffusion and aggregation of signals on graph structures. Defferrard et al. [22] improved spectral graph convolution by using Chebyshev polynomials to achieve fast local filtering of graph signals. Additionally, Veli et al. [23] designed Graph Attention Networks (GAT) to achieve non-uniform aggregation of neighbor node information through adaptively learned attention coefficients, thereby enhancing the model's performance in inductive learning tasks. To address the dependence of traditional models on predefined topological structures, Yu [6] designed Graph Adjacency Learning Networks (GALEN) to automatically construct graph structures through data-driven attention mechanisms. For the synchronous integration of spatio-temporal features, Zhao et al. [24] combined GCN and GRU to construct the T-GCN model, achieving synchronous capture of spatio-temporal correlations. In the specific scenario of abnormal prediction in pumping stations, this paper fuses multimodal information from video and signals and uses GAT to construct a multi-level structural attribute graph, achieving a refined representation of the dynamic dependencies of pumping station operation activities. By introducing a causal dilated convolution module, the model further expands the receptive field exponentially and imposes causal constraints, capturing long-term memory of activity sequences and effectively eliminating the risk of information leakage.

### 2.3 Causal intervention learning

In the process of modeling complex scenarios, shifting deep learning models from simple statistical association learning to causal logic reasoning has become a core path to enhance the reliability and robustness of predictions [25]. Niu et al. [26] utilized causal reasoning theory to transform the problem of "shortcut learning" that models tend to learn, which is highly individualistic and outside the model, into acceptable and interpretable statistical associations. Liu et al. [27] used the backdoor adjustment scheme and inverse probability weighting method to approximately estimate the causal effect of the treatment variable, solving the problem of converting the distribution of causal intervention into the real causal intervention distribution. Liu et al. [28] to compensate for cross-modal bias, calculated the semantic difference between images and text based on the introduction of mutual information theory. Li et al. [29], considering the complex causal relationships in graph-structured data, introduced a spatio-temporal prediction task to identify the true causal effect under the condition that not all confounding factors can be observed, and at the same time, utilized timestamp information to mine key causal subgraphs to act as mediating variables. Yu et al. [30] used the MI-SETC algorithm to enhance a new type of causal decomposition method, successfully constructing stable causal representation means that are insensitive to some disturbances in the entire causal intervention process. Additionally, Huang et al. [31] Adopted the AdaIN technique to achieve feature decoupling, extracting the key causal content from the graph representation from unstable statistical styles. Relevant research demonstrates that causal intervention can effectively guide models to learn the true logical evolution of a system. By doing so, the model is forced to focus on the fundamental initial factors rather than superficial correlations, which significantly improves its generalization performance across different operating conditions and changing scenes.

### 2.4 Summary and motivation

To sum up, although existing research has made progress in spatio-temporal graph neural networks and cross-modal fusion, there are still the following logical gaps: Firstly, most methods assume that the data follows an independent and identically distributed pattern, ignoring the unobserved confounders that are prevalent in industrial environments. Secondly, the existing graph construction methods are mostly static or heuristic, making it difficult to depict the dynamic causal logic evolution during the operation of pumping stations. Based on this, this paper does not merely pursue an increase in model depth but starts from the perspective of causal intervention, aiming to re-examine the generation process of multi-modal features through the front-door adjustment theory. This problem-driven design approach enables the CIFAM model proposed in this paper to fundamentally block spurious correlations, thereby demonstrating stronger physical consistency when dealing with asynchronous time-lagged signals.

# 3 Construction of Structural Attribute Graph

In response to the insufficiency of capturing the logical flow of the pump station operation process in the previously mentioned method, this chapter mainly explains how the structure attribute graph in the CIFAM model is established. By drawing on the interpretability advantage of the attention mechanism and referring to the approach in Reference [6], this paper proposes an automatic graph construction method based on whether there is a connection between activities. If there is a strong connection, a link is established between the two activity points to capture the associations among various activities during the pump station operation process. Figure 1 illustrates the process of graph construction.

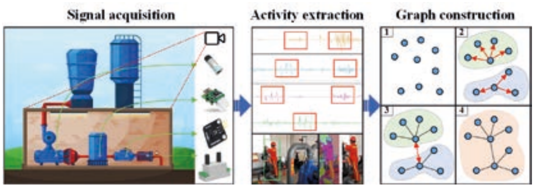


**Figure 1**. Graph construction process

In the data preprocessing and feature modeling stage, for the video stream and high-frequency sensor signals with significant differences in sampling frequencies, this paper adopts a synchronization mechanism based on UTC timestamps [32]. By setting a sliding window of length $W$ and downsampling the sensor data to match the video frame rate, it ensures the precise alignment of multimodal information under a unified temporal reference. At the feature extraction level, this paper utilizes a pre-trained ResNet-50 network to extract the spatial semantics of video frames, and employs a three-layer 1D-CNN to capture the local temporal correlations of sensor signals. Further, by integrating temporal action localization and change point detection algorithms [33, 34], it automatically segments and defines the semantic segments of operational behaviors in the video and the operational phases of the signals, respectively. Ultimately, the multi-source heterogeneous features are mapped to a unified embedding space, serving as the initial node representations of the structural attribute graph. To optimize the aggregation process of neighborhood information, a graph attention network (GAT) [23] is employed to introduce an adaptive learning mechanism, which enhances the accuracy of feature extraction by automatically computing the attention weights of adjacent nodes. In a given graph $G(V,E)$ , the attention coefficient for $edge(i,j)$ measures the interaction strength between nodes as:

$$e_{ij} = \mathrm{LeakyReLU}\left(\mathbf{a}^{\top}[\mathbf{W}\mathbf{h}_i \,\|\, \mathbf{W}\mathbf{h}_j]\right), \quad (i,j) \in E \tag{1}$$

Here, $\vec{h}_i$ and $\vec{h}_j$ represent the features of nodes $v_i$ and $v_j$ respectively. The model uses the weight matrix $W$ to perform embedding mapping on the node features and combines the attention vector $\vec{a}$ to evaluate the degree of correlation between the two feature embeddings. To ensure numerical stability and facilitate cross-node comparisons, the model ultimately normalizes the calculated attention coefficients using the softmax function to obtain the final weight distribution, as shown in the following formula:

$$a_{ij} = \mathrm{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ij})} \tag{2}$$

Among them, $N_i = \{k \in N : (i,k) \in E\} \cup \{i\}$ is the self-contained adjacency set of node $V_i$. The new feature $\vec{h}'_i$ of node $V_i$ can be aggregated in the following way:

$$\vec{h}'_i = \sigma\left(\sum_{j \in N_i} a_{ij} W \vec{h}_j\right) \tag{3}$$

In the formula, $\sigma(\cdot)$is processed using the ReLU activation function. Due to the high correlation between the topological features of the graph and the fault evolution pattern, strongly associated node pairs are regarded as the basis for the existence of edges in the graph. By sorting all attention coefficients $a_{ij}$ in descending order and selecting the top $D \times N$ connections with the highest scores, the edge set $E$ is constructed. Here, the parameter $D$ is used to flexibly adjust the average degree of nodes, thereby extracting intuitive structural information $S$ from the activity association graph $G(V,E)$.

According to previous studies, moderately contracting the receptive field of the attention mechanism can enhance the model's ability to capture core dependencies. Based on this principle, the complex pumping station activity process is divided into $K = \sqrt{N}$ different clusters using the K-Means [35] algorithm. Then, within each cluster group, the aforementioned weight calculation method is applied to determine the connection attributes between nodes. Finally, the same association calculation is performed for the $K$ central nodes to complete the construction of the entire multi-level structure attribute graph $G$.

Although the multi-level structure attribute graph $G$ provides a foundation for representing operational dependency relationships, it is still susceptible to unobserved confounding factors. Therefore, the next chapter will introduce a causal intervention framework to further eliminate these spurious correlations.

## 4 Multi Modal Causal Intervention

In the task of pump station anomaly prediction, we focus on identifying the true causal effect of the multi-modal structural attribute graph $G$ on the final prediction result $Y$, denoted as $P(Y|do(G))$. However, as illustrated in the structural causal model (Figure 2), confounding factors $B$ (e.g., erratic environmental lighting or latent operational variances) introduce spurious correlations that render traditional back-door adjustment inapplicable. To resolve this, we invoke the Front-door Criterion by identifying a mediator variable that intercepts the causal path from $G$ to $Y$.

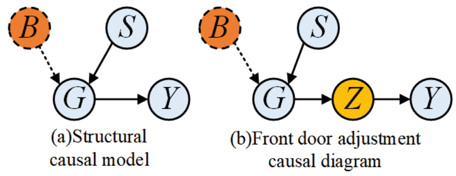


**Figure 2**. Causal structure model

Theoretical Identification: A variable $Z$ is said to satisfy the front-door criterion if: (i) intercepts all directed paths from $G$ to $Y$; (ii) there is no unblocked back-door path from $G$ to $Z$; and (iii) all back-door paths from $Z$ to $Y$ are blocked by $G$. In CIFAM, we treat the extracted invariant causal subgraph as a principled heuristic approximation of the mediator $Z$. By optimizing a contrastive loss under the Causal Invariance Principle, we force $Z$ to capture representations that remain robust across local topological perturbations (simulated by random edge dropping), thereby effectively blocking the interference of unobserved confounders and aligning the neural architecture with the probabilistic intent of the front-door adjustment formula:

$$P(Y|do(G)) = \sum_{z} P(z|G) \sum_{g} P(Y|g, z)P(g) \quad (4)$$

Where $z$ represents the realizations of the mediator $Z$, and $g$ is a summation variable representing the space of all possible structural graph configurations $G$. Under the assumption that causal structures remain stable while spurious correlations are fragile, this mechanism effectively blocks interference from $B$. Specifically, the random edge dropping mimics the intervention on $G$, ensuring that the identified representation $Z$ is driven by the internal structural properties of the graph rather than being misled by unobserved confounding correlations, thereby aligning with Condition (ii).

Based on the above core causal mechanism, this paper proposes the first multimodal causal intervention model CIFAM that effectively incorporates the front-door adjustment strategy into the prediction of pump station anomalies. The architecture diagram of CIFAM is shown in Figure 3. Firstly, CIFAM builds the structural attribute graphs of video and signal using different modal information respectively, and uses the graph structure information $S$ to enhance and simulate the causal intervention on $G$, generating the augmented graph $\widetilde{G}$. Then, through contrastive learning, it identifies and learns the in-

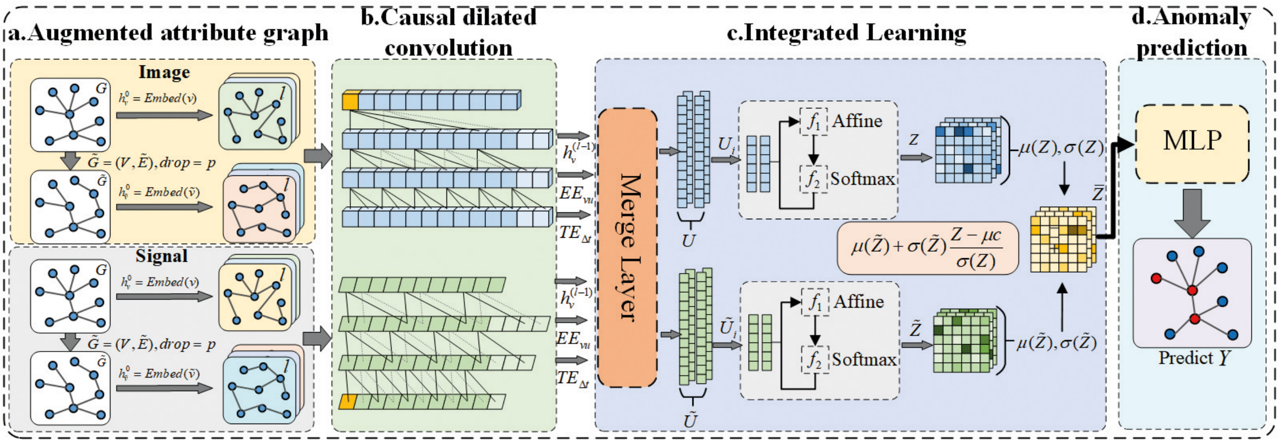


**Figure 3**. Architecture diagram of the CIFAM Model.

termediate representations insensitive to the intervention from $G$ and $\widetilde{G}$, that is, the representations of the causal subgraphs, and finally uses this intermediate representation to predict the abnormal result $Y$.

This section will provide a detailed introduction to the core mechanisms of the model, mainly including graph structure enhancement, multi-level representation encoding, representation fusion based on adaptive instance normalization, and anomaly predictor. The core innovation of this paper lies in the first-time introduction of the front-door adjustment idea into the pump station anomaly prediction task, and for this purpose, a causal intervention learning framework based on front-door adjustment is designed.

## 4.1 Overview of data flow

The implementation of the front-door adjustment in CIFAM is realized through a sequential data flow involving four primary modules. To clarify the correspondence between the mathematical formulations and the model structure, the input-output relationships are defined as follows:

Augmented Graph Module: Takes the initial structural attribute graph $G(V,E)$ as input and outputs the augmented graph $\widetilde{G}(V,\widetilde{E})$ (Equation 4), simulating the causal intervention process.

Representation Encoding Module: Inputs $G$ and $\widetilde{G}$ into the causal dilated convolution and multi-head attention to capture spatio-temporal dependencies, outputting layer-wise node representations $h_v^{(l)}$.

Causal Fusion Module (AdaIN): Aggregates node features into graph-level representations $Z$ and $\widetilde{Z}$, then executes the front-door adjustment via Adaptive Instance Normalization to output the deconfounded representation $\bar{Z}$.

Prediction Module: Takes $\bar{Z}$ as the input to a Multi-Layer Perceptron to generate the final anomaly prediction $\widehat{Y}$.

This design of sequential data flow is essentially a neural implementation of the front-door adjustment criterion.

## 4.2 Augmented graph generation

Pump station data may be affected by various noises and uncertainties in the real transmission process, such as unexpected interruptions of certain activities or incomplete observations. To enable the model to learn representations that can resist these external perturbation factors and focus on the abnormal factors of the activities themselves, CIFAM introduces graph structure augmentation as an intervention measure. To simulate causal intervention and identify robust activity anomaly propagation paths during operation, CIFAM introduces graph structure augmentation. Specifically, based on the principle of causal invariance, a true causal structure, such as a key propagation path, should have a certain degree of robustness against random, local perturbations in the network, while spurious correlation often rely on specific, unstable shortcut paths that are easily disrupted under intervention. Therefore, for each input original structural attribute graph $G(V,E)$, CIFAM adopts a random edge dropping strategy to generate an augmented view $\widetilde{G}(V,\widetilde{E})$. Formally, the generation process of the augmented edge set $\widetilde{E}$ is defined as follows:

$$\widetilde{E} = \{e_{ij} \in E | m_{ij} = 1\} \tag{5}$$

Among them, $m_{ij}$ is the mask variable corresponding to edge $e_{ij}$, which follows a Bernoulli distribution with parameter $1-p$:

$$m_{ij} = \text{Bernoulli}(1\text{-}p) \tag{6}$$

In this formula, $p \in [0,1]$ represents the preset dropout rate. This operation constructs counterfactual samples by randomly perturbing the topological structure while keeping the node features $h_v^0$ unchanged. This intervention on the graph structure can be regarded as generating counterfactual samples on the original observed data, thereby reducing the model's over-reliance on certain specific edges during the prediction process, preventing the model from learning spurious correlation dominated by only a few paths, and effectively guiding the model to prefer learning causal relationships rather than spurious correlation, thereby improving the model's generalization ability. Although the random edge deletion strategy provides effective causal intervention for most industrial topological structures, its reliability is inherently related to the density of the graph. In extremely sparse structures, purely random perturbations may inadvertently remove key causal path edges. To alleviate this failure mode, CIFAM maintains a conservative discard rate $p$ and relies on a multi-view contrast mechanism to ensure that key causal information can still be identified even under structural perturbations.

Meanwhile, CIFAM adopts a contrastive learning framework, which forces the model to focus on the more essential structures that exist in both views, learning a representation that is invariant to perturbations rather than relying on a few edges that may be randomly dropped. Additionally, for the same structural attribute graph throughout the training process, the augmented views generated in each iteration are different. By training under a large number of random perturbations, the model can statistically learn which structures persist and which occur by chance. This effectively mitigates the problem of real correlation loss that may be caused by random edge dropping.

To learn how information affects the most abnormal prediction results from the complex pump station activity structure and temporal dynamic, and to effectively integrate information from the original structure attribute graph $G$ and the augmented structure attribute graph $\widetilde{G}$, CIFAM designs a mechanism that includes multi-layer representation encoding and cross-graph representation fusion. This mechanism first independently generates graph-level representations for $G$ and $\widetilde{G}$, and then fuses them through adaptive instance normalization.

For a given structure, the node representation is learned through a recursive, layer-by-layer aggregation of neighbor information process. Let $h_v^{(l)}$ represent the representation of node $v$ at layer $l$. The initial representation of the node is composed of its inherent sequence embedding:

$$h_v^0 = NodeEmbed(v) \tag{7}$$

Among them, $NodeEmbed$ is a learnable node embedding lookup function. Subsequently, at each layer $l$, the representation $h_v^{(l)}$ of node $v$ is updated and aggregated based on its own representation $h_v^{(l-1)}$ in the previous layer and the information of its temporal neighbor node set $N(v,t_j)$, where $N(v,t_j)$ is the set of neighbor nodes of $v$ before the current observation cutoff time $t_j < t_o$.

## 4.3 Causal dilated convolution

Before applying the graph attention mechanism for spatial topological aggregation, CIFAM deeply models the temporal features of nodes and their neighbors through the causal dilated convolution module. The core purpose of introducing the causal dilated convolution module is to capture the long-term memory of the activity sequence by obtaining an exponentially expanded receptive field, while ensuring that the model does not rely on future moment data through causal constraints, thereby eliminating the risk of “information leakage”. For the neighbor node set $N(v,t_j)$ of node $v$ before the observation moment $t_j$, its representation sequence at the $l-1$ layer is denoted as:

$$X_v = \{h_{u_1}^{(l-1)}, h_{u_2}^{(l-1)}, \ldots, h_{u_n}^{(l-1)}\} \tag{8}$$

The enhancement process of the time series features by using the filter $f:\{0,\ldots,k-1\}\to R$ and the dilated causal convolution operator $\otimes d$ is defined as follows:

$$\widehat{h}_u^{(l)}(s)=(X_v\otimes d)(s)=\sum_{i=0}^{k-1}f(i)X_{v_s-d_i} \tag{9}$$

Among them, $d$ is the dilation coefficient, which allows the model to obtain historical information over a longer time step without significantly increasing the computational cost during the exponential growth process with the network depth. $k$ represents the size of the convolution kernel, characterizing the width of the window for local feature extraction. $v_s-d_i$ embodies the causal constraint, by padding zero elements at the front end of the sequence and restricting the convolution index, ensuring that the output at time S is calculated only using the current and historical states, strictly preventing the leakage of future information to the past.

The causal dilated convolution module plays a crucial role in preprocessing and feature enhancement in the model proposed in this paper. In complex pumping station operating conditions, abnormal signals often have significant time lags. The large receptive field of the causal dilated convolution module enables the model to trace back to the operation parameters at earlier times and extract more representative evolution features. Moreover, by eliminating information leakage, the model can learn the interaction logic between nodes based on the true causal order, thereby reducing the spurious correlation caused by unobservable variables.

## 4.4 Graph fusion learning

After processing through the causal dilated convolution module, the information representation passed from neighbor $u$ to $v$ is used as the $k_{vu}^{(l)}$ key and $v_{vu}^{(l)}$ value in the attention mechanism, which can be concatenated as follows:

$$k_{vu}^{(l)}=v_{vu}^{(l)}=\mathrm{Concat}(\widehat{h}_u^{(l)},\mathrm{EE}_{vu},\mathrm{TE}_{\Delta t}) \tag{10}$$

$$\mathrm{EE}_{vu}=edgeEmbed(e_{vu}) \tag{11}$$

$$\Delta t=t_j-t_{vu} \tag{12}$$

$$\mathrm{TE}_{\Delta t}=TimeEncoder(\Delta t) \tag{13}$$

Here, $\mathrm{EE}_{vu}$ represents the type embedding of edge $e_{vu}$, and $edgeEmbed(\cdot)$is the type embedding function. $\Delta t$ denotes the difference between the current observation time $t_j$ and the historical interaction time $t_{vu}$ between nodes $v$ and $u$. After encoding through the time encoding function $TimeEncoder(\cdot)$, the relative time difference encoding $\mathrm{TE}_{\Delta t}$ is obtained. Concat indicates the concatenation operation.

The central node $v$ generates the corresponding query vector $q_v^{(l)}$ at the $l$ layer, whose construction method is similar to that of keys and values, usually composed of its representation from the previous layer and its time information:

$$q_v^{(l)}=\mathrm{Concat}(\widehat{h}_v^{(l-1)},\mathrm{EE}_v,\mathrm{TE}_{\Delta t=0}) \tag{14}$$

Here, $EE_v$ represents a placeholder edge embedding designed for the center node $v$ itself, which is a special, learnable vector and will be learned and updated like other edge embeddings during the training process. $\mathrm{TE}_{\Delta t=0}$ represents the zero-time-difference time encoding, which is a tensor of all zeros. The key $k_{vu}^{(l)}$, value $v_{vu}^{(l)}$, and query vector $q_v^{(l)}$ are input into the multi-head attention module together and fused with the node representation of the previous layer to generate the representation $\widehat{h}_v^{(l)}$ of the $l$ layer:

$$Aggre_v^{(l)}=Multi(q_v^{(l)},\{k_{vu}^{(l)}\},\{v_{vu}^{(l)}\}),u\in N(v,t) \tag{15}$$

$$\widehat{h}_v^{(l)}=Merge(Aggre_v^{(l)},\widehat{h}_v^{(l-1)}) \tag{16}$$

Here, $Multi(\cdot)$represents the multi-head attention module, and $Merge(\cdot)$is the concatenation layer. The two node representation vectors are concatenated and then sent into a Multi-Layer Perceptron, (MLP) for nonlinear transformation, ultimately obtaining the representation $h_v^{(l)}$ of the $l$ layer.

In CIFAM, the above layer-by-layer node representation learning process is respectively applied to $G$ and $\widetilde{G}$. Next, for each graph, the model performs layer-by-layer aggregation on the source node and target node lists, thereby obtaining node representations of different depths. During the layer-by-layer fusion, since the node representations of the first layer undergo a single neighbor aggregation, they mainly capture the local, fine-grained structural features of the structural attribute graph, such as direct activity dependencies and early anomaly propagation patterns. In contrast, the node representations of the last layer undergo multiple neighbor aggregations, and the receptive field of the nodes has expanded to the entire observed graph. Therefore, the representations of this layer can capture the global, high-order topological structure of the structural attribute graph. Thus, CIFAM chooses to fuse the node representations of the first layer and the last layer to generate the graph-level intermediate representation:

$$U^{(l)} = Merge(\mathrm{Pool}(H^{(l)}_{v\in S}), \mathrm{Pool}(H^{(l)}_{v\in T})) \quad (17)$$

Here, $l \in \{0, -1\}$ represents the layer number, which can be either the first layer or the last layer. $H^{(l)}_{v\in S}$ and $H^{(l)}_{v\in T}$ respectively denote the source node and target node representations of a specific layer, and Pool$(\cdot)$represents the pooling operation. By fusing these two representations at different scales, CIFAM can simultaneously utilize micro and macro structures for prediction, which helps prevent the common "over-smoothing" problem in deep graph neural networks, where deep node representations tend to homogenize, leading to the loss of local information and thus retaining the multi-scale features crucial for anomaly prediction.

For each fused intermediate graph representation $H^{(l)}_{v\in T}$, the model calculates its importance weight $a^{(l)}$ through an affine transformation and a Softmax function, and then performs a weighted sum to obtain the pooling vector of this fused intermediate graph representation:

$$a^{(l)} = \mathrm{Softmax}(\mathrm{Affine}(U^{(l)})) \quad (18)$$

$$\widehat{u}^{(l)} = \sum_{nodes} (U^{(l)}, a^{(l)}) \quad (19)$$

Ultimately, all the selected and weighted pooled graph representations are collected and weighted summed through a learnable layer weight vector $w_{layer}$ to obtain the final causal intervention graph representation $Z$:

$$Z = \sum_{i} w_{layer} \cdot \widehat{u}^{(l)} \quad (20)$$

After obtaining the graph-level representations of the original graph $Z$ and the augmented graph $\widetilde{Z}$ after cross-modal fusion, to integrate the learned information from the two views and generate a more robust final representation that can suppress spurious correlation related to partial and confusing factors, the model introduces an Adaptive Instance Normalization (AdaIN) [31] module for feature fusion:

$$\bar{Z} = AdaIN(Z, \widetilde{Z}) = \mu(\widetilde{Z}) + \sigma(\widetilde{Z})\frac{Z - \mu(Z)}{\sigma(Z)} \quad (21)$$

Here, $\mu(\cdot)$and $\sigma(\cdot)$respectively represent the mean and standard deviation, both of which are calculated along the feature dimension. The core design principle lies in decoupling the graph representation into two parts: "true factors" and "confounding factors". In this task, true factors can be understood as the stable, causal core structure that is robust in the process of causal intervention. This is the part that the model is expected to learn and use for prediction. Confounding factors can be regarded as unstable statistical features brought about by random disturbances in complex scenarios, non-critical edges, etc., and this part is very likely to contain spurious correlations.

AdaIN achieves feature transformation by aligning the real factors of $Z$ with the confounding factors of $\widetilde{Z}$. Here, CIFAM assumes that the statistical features of $\widetilde{Z}$, after graph augmentation intervention, become more robust due to the disruption of some fragile connections, thus retaining the content of the original structural attribute graph representation $Z$ while replacing its confounding factors with the real factors of $\widetilde{Z}$.

This operation can be regarded as a kind of denoising of the original graph representation, aiming to eliminate its potentially unstable statistical characteristics caused by spurious correlation, thereby extracting a final representation $\widetilde{Z}$ that is closer to the causal essence. Mathematically, the optimization of the contrastive loss serves as the numerical engine for identifying the mediator $Z$ required by the front-door criterion. By maximizing the mutual information between the treatment $G$ and its augmented view $\widetilde{G}$, the contrastive objective forces $Z$ to capture representations that remain invariant across local topological perturbations. In the context of the front-door adjustment, this process ensures the identifiability of the conditional probability $P(z|do(G))$, while the AdaIN based feature decoupling effectively approximates the summation over the state space of $G$ in Equation (4). Consequently, the minimization of the contrastive loss drives the neural architecture to converge toward the theoretical solution of the front-door formula, ensuring that the model's predictions are rooted in stable causal logic rather than transient shortcut features. Ultimately, the graph representation $\bar{Z}$ obtained through AdaIN fusion is sent to a predictor for the final abnormal activity prediction.

### 4.5 Anomaly predictor

The predictor employs a Multi-Layer Perceptron (MLP) composed of two fully connected layers (with ReLU activation functions) to map the fused graph representation to the predicted value $\widehat{Y}$:

$$\widehat{Y}=\mathrm{MLP}(\bar{Z}) \tag{22}$$

The training objective of the model is to minimize the deviation between the predicted value and the true value. To alleviate the common confusion factors in the labels of real pumping station anomaly detection and enhance learning stability, the model does not directly predict the abnormal value $\Delta Y$, but rather its logarithmically transformed value $\widehat{Y}$. The preprocessing process is defined as follows:

$$\widetilde{Y} = \log_2(\Delta Y + 1) \tag{23}$$

The model employs mean squared error and hyperparameters as the loss function:

$$Loss = \frac{\lambda}{N}\sum_{i=1}^{N}(\widehat{Y}_i - \widetilde{Y}_i)^2 \tag{24}$$

Here, $N$ represents the batch size, and parameter $\lambda \in [0,1]$ is the causal intervention degree adjustment factor introduced in this paper. By minimizing this loss function, the model continuously optimizes its parameters, thereby enhancing the accuracy of anomaly prediction.

## 5 Experiment and Performance Analysis

This section conducts a systematic experimental evaluation of the proposed CIFAM model's effectiveness and robustness in complex industrial environments. By analyzing key performance indicators, it demonstrates the model's effectiveness in capturing real causal logic and suppressing spurious correlation in various non-steady-state scenarios.

### 5.1 Dataset and methodology

To further demonstrate the effectiveness and generalization ability of the CIFAM model proposed in this paper for non-intrusive anomaly detection, verification was conducted using five common public benchmark datasets from the industrial field and a self-made actual pump station non-intrusive monitoring dataset. The public datasets include MVTecAD, Real-IAD, MulSen-AD, MPDD, and CAD-SD, covering various complex industrial manufacturing and detection scenarios, which can better provide standardized evaluation references for model performance. Meanwhile, We have constructed a multimodal real-world dataset for non-intrusive monitoring of pumping stations (our dataset), which was collected during the operation of a real pumping station. This dataset includes real-time images from 1148 video streams as well as various operational conditions such as current, vibration, temperature, and noise. The eight different clusters mentioned in the experimental analysis correspond to specific physical states: (1) normal operation, (2) impeller blockage, (3) bearing wear, (4) cavitation, (5) seal leakage, (6) motor overload, (7) misalignment, and (8) lubrication failure. The real anomaly labels were generated through a strict semi-automatic process; initially, the initial anomaly boundaries were marked using preset hardware thresholds, followed by manual verification and semantic annotation to ensure that the labels were consistent with actual operational events, the final dataset presents a balanced distribution. The sampling rate of the vibration sensor is 10 kHz, the sampling rate of the current and noise sensors is 1 kHz, and the frame rate of the 1080p video stream is 25 frames per second. All modalities are aligned using UTC.

Baseline models and the CIFAM model output the accuracy and Micro-F1 score of the next moment in the test set. All models were trained using the Adam optimizer with a learning rate of 0.001. 60% of the nodes were used as the training set, and the remaining 20% of the data were respectively used as the validation set and the test set. When the dynamic graph structure changes, the baseline models cannot update the model based on the graph topology structure. Therefore, in the baseline methods, the input is the node features at the current moment and the graph structure information at the initial moment; in the CIFAM model, the input is the information at the current moment.

In terms of experimental setup, this paper conducts semi-supervised node anomaly detection tasks on all the aforementioned datasets. The main baseline models examined in the experiments are:, EGNN [36], AGP [37], DySAT [38], EvolveGCN [39], SEIGN [40], ContinualGNN [41], SDGNN [42] and SpikeNet [43]. To highlight the advantages of causal reasoning, we added two state-of-the-art causal baseline models: CIGNN[4], which utilizes causal intervention for industrial fault diagnosis, and CUBE [44], a framework based on counterfactual explanations to mitigate spurious correlations. During the model training phase, only the activity data under normal operating conditions are used to construct the initial graph structure, aiming to capture the normal spatiotemporal evolution patterns through an adaptive dynamic graph learning mechanism, thereby effectively identifying abnormal activity nodes that deviate from the center of the normal distribution in the testing phase.

## 5.2 Evaluation indicators

To verify the effectiveness and stability of the proposed CIFAM for the task of lossless anomaly detection, node classification Accuracy and Micro-F1 score are adopted for comparative analysis here. Accuracy is a commonly used intuitive evaluation metric that can measure whether the model correctly classifies all nodes of normal and abnormal activities. The calculation formula is shown in Equation (25).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{25}$$

Among them, $TP$ represents the number of abnormal nodes correctly identified, $TN$ represents the number of normal nodes correctly identified, $FP$ represents the number of normal nodes wrongly identified as abnormal, and $FN$ represents the number of abnormal nodes wrongly identified as normal. Due to the possible imbalance in sample distribution in industrial anomaly detection scenarios, the simple accuracy rate may not fully reflect the model performance. This paper adopts the Micro-F1 score to comprehensively measure the Precision and Recall of the model. Micro-F1 calculates the global indicator by summarizing the total true positives, false positives, and false negatives of all categories, and is particularly suitable for multi-classification problems. First, calculate the global precision $Micro-P$ and global recall $Micro-P$ as shown in formulas (26) - (27).

$$Micro-P = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i+FP_i)} \tag{26}$$

$$Micro-R = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i+FN_i)} \tag{27}$$

Among them, $N$ represents the total number of categories (corresponding to the 14 activity types defined in this article), and $TP_i$, $FP$, and $FN_i$ respectively represent the number of true positives, false positives, and false negatives of the second category. The calculation formula for the Micro-F1 score is.

$$Micro-F1 = \frac{2\cdot Micro-P\cdot Micro-R}{Micro-P+Micro-R} \tag{28}$$

## 5.3 Abnormal prediction result

From the Micro-F1 values of different methods in Figure 4, it can be seen that the CIFAM proposed in this paper has certain advantages over competing models. By comparing the data in Figure 4, it can be observed that CIFAM has achieved the best prediction results, with significant improvements in the Micro-F1 metric compared to traditional topological models such as Graph SAGE and DySAT, as well as recent causal diagnosis frameworks like CUBE and CIGNN. In terms of generalization ability, CIFAM demonstrates excellent performance across various types of datasets, especially with a prediction accuracy of 91.1% on the OURS dataset. This is due to its stronger ability to integrate multimodal heterogeneous signals and identify true causal evolution paths. As expected, the causal-based baselines demonstrate superior performance compared to non-causal models, validating the importance of causal reasoning in industrial scenarios. In contrast, although the overall prediction scores of the CAD-SD dataset are lower than other benchmarks due to its extreme noise, CIFAM still achieves the highest Micro-F1 score (59.7%) among all tested models.

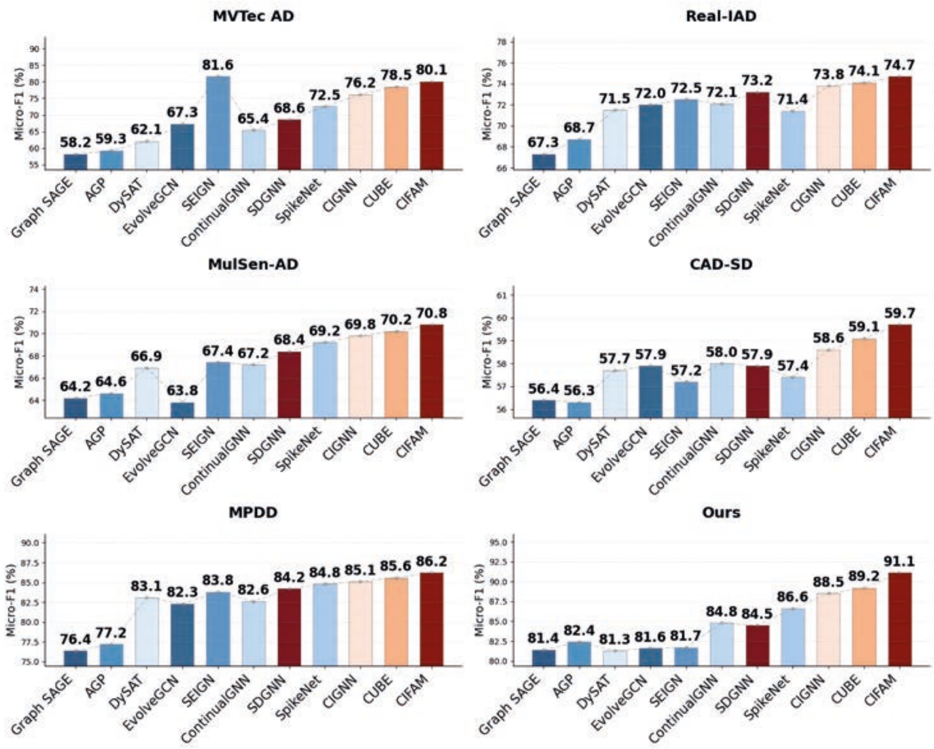


**Figure 4**. Comparison of Accuracy by different methods

## 5.4 Convergence and stability analysis of multi-class models

To comprehensively evaluate the learning efficiency and numerical stability of the CIFAM model in multi-objective optimization, this section compares the training convergence curves of CIFAM with three representative baseline models on the OURS dataset. SpikeNet representing high-performance deep feature extraction; EvolveGCN representing mainstream spatio-temporal evolution modeling; DySAT representing complex self-attention-based dynamic graph architectures; CIGNN representing state-of-the-art causal intervention logic. The comparison models include high-performance feature extraction model SpikeNet, mainstream spatio-temporal models EvolveGCN and DySAT, and the state-of-the-art causal model CIGNN.

As shown in Figure 5, in terms of convergence speed and quality, all models can rapidly converge to relatively low loss values in the initial stage. However, due to their inability to capture complex nonlinear spatio-temporal dependencies or causal relationships, they eventually get stuck at a relatively high loss value, that is, they fall into local optima. In terms of training stability, DySAT experiences some fluctuations during training due to the use of a complex self-attention mechanism. In contrast, although CIFAM's learning curve is slightly delayed in the early stage due to the addition of causal intervention, it can still maintain a relatively smooth state in the middle and later stages, and CIFAM can achieve the lowest loss value throughout the entire training process among all models, indicating that it does not sacrifice convergence for accuracy. Meanwhile, by filtering out some noise through the front-door adjustment method, the features learned by CIFAM can more directionally capture the incremental gains brought by the causal mechanism, thereby also giving it better training stability than traditional dynamic graph models.

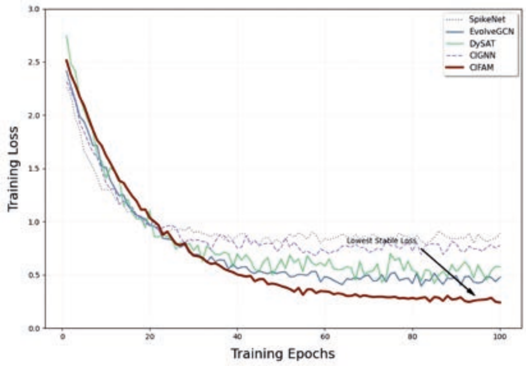


**Figure 5**. Training convergence and stability

## 5.5 Ablation study

To verify the role of each important module of the CIFAM model in enhancing the performance of anomaly detection and to explore whether the performance gains are truly due to causal factors, ablation experiments were conducted on four representative datasets: Real-IAD, MulSen-AD, MPDD, and Ours. Different variant models were compared, and the performance results of these variants were used to determine whether the causal step, sequence modeling, and feature fusion modules could be confirmed to have a significant impact on the model's predictive ability.

Removing the causal dilated convolution module (w/o Convolution): This variant removed the causal dilated convolution module mentioned above and used either ordinary convolution or directly used the original time series features for spatial topology aggregation. Due to the loss of the advantage of exponentially expanding the receptive field through the dilation coefficient, it was unable to effectively track the operation parameters and states at earlier time points when dealing with pump station anomaly signals with strong time lags.

Removing the attention module (w/o Attention): The attention module was removed and replaced with simple mean pooling to determine its role in aggregating neighbor information and selecting key nodes.

Removing causal intervention (w/o Intervention): The graph data augmentation module was removed, which means the core causal intervention method was abandoned. At this point, the model degraded to a simple supervised learning approach, only able to learn some related factors from the observed data but unable to clearly distinguish between cause and confusion.

Removing causal fusion (w/o Fusion): This variant deleted the AdaIN fusion module. Since the model did not align or fuse the representations learned before and after the intervention causally, after removing causal fusion, the model also lacked the ability to filter and retain invariant features under interference.

As shown in Table 1, in the w/o Convolution variant, the model degrades to a simple convolution or direct spatial aggregation of the original time series information. Due to the loss of the ability to exponentially increase the receptive field through the expansion coefficient, the model is unable to effectively trace back the operation parameters at distant times for the significant time-lagged characteristics of the pump station's abnormal signals. On the Ours dataset, the performance drops from 91.1% to 87.6%. By removing the w/o Attention model and replacing the attention mechanism with simple mean pooling, we aim to demonstrate the model's effectiveness in aggregating neighbor information and identifying key nodes. The results show that non-adaptive weight allocation leads to the model's inability to accurately model the correlations between complex activities. After removing the w/o Intervention module, the model reverts to a traditional supervised learning mode. In this case, the model directly learns statistical correlations from the observed data rather than causal or confounding factors, making it difficult to prevent "spurious correlation" caused by environmental noise. After removing the AdaIN fusion module, although the model performs intervention, it fails to align the representations before and after the intervention. As a result, the model cannot extract the core causal content with strong robustness from the disturbances and cannot achieve feature decoupling.

From the experimental data, it is evident that causal dilated convolution has the most significant impact on the model's performance. In the Ours pump station dataset, it causes the most severe performance degradation, indicating that in actual complex industrial time series scenarios, only by establishing long-term temporal dependencies can accurate predictions be made. Moreover, combining causal intervention and fusion mechanisms can probabilistically enhance the model's robust-

**Table 1**. Ablation experiments on four datasets

| **Data Set/Module** | Real-IAD | MulSen-AD | MPDD | Ours |
|---|---|---|---|---|
| w/o Convolution | 69.7/66.4 | 67.4/62.1 | 76.7/71.5 | 87.6/83.8 |
| w/o Attention | 71.2/71.9 | 67.8/64.7 | 77.2/73.5 | 88.9/88.3 |
| w/o Intervention | 73.8/73.3 | 69.5/67.8 | 83.1/82.3 | 90.3/89.8 |
| w/o Fusion | 73.8/73.3 | 69.5/67.8 | 82.3/78.6 | 90.3/89.8 |
| CIFAM | 74.7/74.2 | 70.8/69.1 | 83.8/82.4 | 91.1/90.6 |

ness to unseen confounding factors and ensure that the model's predictions are based on physically realistic logic rather than shortcut learning.

### 5.6 Parameter sensitivity analysis

The sensitivity of the CIFAM model to key hyperparameters and its stability in actual training were analyzed in this paper. The intervention intensity $\lambda$ and the causal edge dropping probability $p$ were set as the main factors affecting CIFAM. Multiple rounds of experiments were conducted here for verification. For all parameter configurations set, the model was run five times on the OURS dataset separately, and the mean and variance of the Micro-F1 metric in each experiment were recorded. The experimental results are detailed in Figure 6, where the solid line represents the average of the five experiments, and the shaded area represents the variance range. The causal intervention intensity $\lambda$ determines the degree of attention the model pays to the extraction of causal mediator representations during the training process. Figure 6(a) shows that as $\lambda$ gradually increases, the average value of the model first rises and then falls; when $\lambda = 0.5$, the model's prediction effect is the best and the variance range is the narrowest, indicating that implementing causal intervention at $\lambda = 0.5$ can avoid interference factors to the greatest extent. Additionally, When the value of A is too high or too low, the range of variation will increase, indicating that maintaining a specific balance is crucial for ensuring the stability of the model. The causal edge dropping probability $p$ is used in the training stage to represent invisible environmental disturbances by randomly deleting graph edges. The results shown in Figure 6(b) indicate that when $p$ is in the range of [0.2, 0.4], the model's performance is stable and the fluctuation range is small, suggesting that the front-door optimization method, by randomly deleting edges to simulate environmental disturbances, enables the model to learn causal mediator representations that are less sensitive to local topological perturbations. Although CIFAM performs stably within the range of $p$ in [0.2, 0.4], its performance drops significantly when $p$ exceeds 0.5. This trend is further clarified at the boundary of the high-performance platform in the heatmap. This performance decline is mainly due to the excessive deletion of sparse causal edges. In some pump station sections, operational activities are strictly sequential and without redundancy, and excessive intervention intensity can disrupt the critical causal logic flow. Therefore, the determined optimal parameter range reflects a balance between effectively blocking confounding paths and preserving the key structural backbone.

In addition to the single-parameter cross-experiment exploring the co-evolution of causal intervention intensity $\lambda$ and intervention probability $p$, in this section, we also conducted a two-parameter cross-experiment based on the OURS dataset, with different value ranges of parameters $\lambda$ and $p$. We performed a grid search with $\lambda$ and $p$ taking values in the intervals [0.1, 0.9] and [0.1, 0.8] respectively, and recorded the corresponding Micro-F1 values for each pair of different values. The heat map is shown in Figure 7. As shown in Figure 7, when $\lambda$ is within the range of [0.4, 0.6] and $p$ is within [0.2, 0.4], the model has a high-performance plateau region, indicated by the dark area, where the Micro-F1 is all above 90%. At this point, it is a better choice to provide sufficient perturbation samples for the front-door adjustment by setting a reasonable intervention probability and to cooperate with a certain intensity of causal constraint conditions to exert a greater extraction effect of the mediating effect. If both are at a high or low level, it indicates that too many loss features or there is a large statistical deviation, and there will

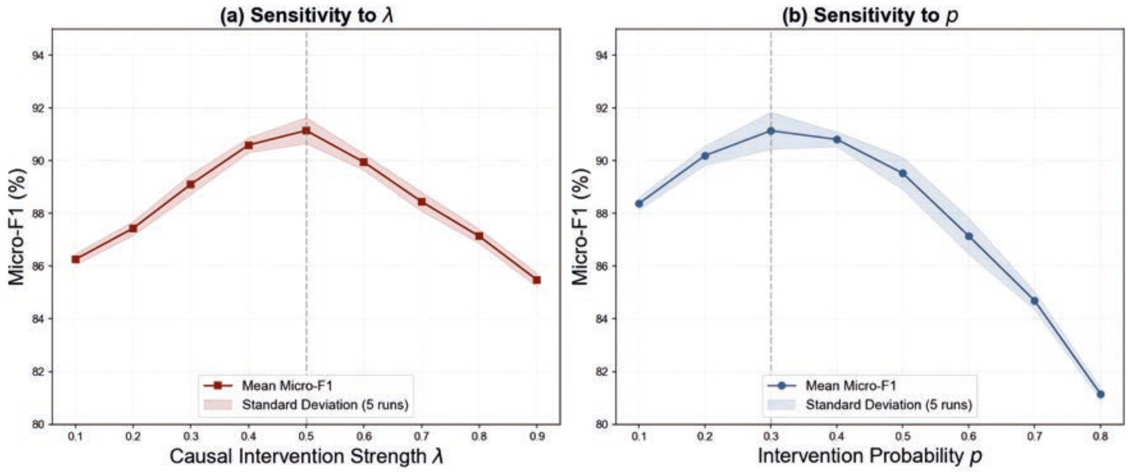


**Figure 6**. Results of hyperparameter sensitivity analysis

be a situation where one of them increases and the other necessarily decreases. Besides finding the optimal parameters of the model, it also proves the stable performance of the CIFAM model under different parameter couplings.

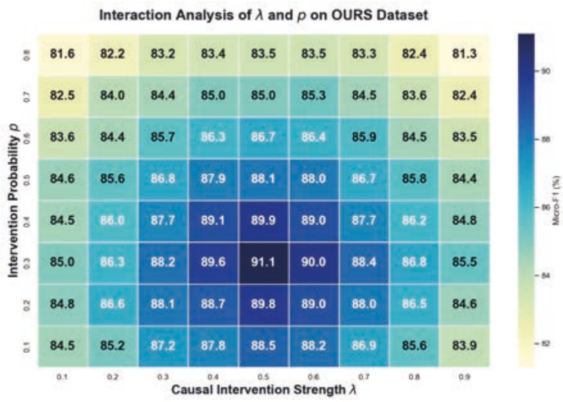


**Figure 7**. Heatmap of interactions among hyperparameters

## 5.7 Effectiveness analysis of causal representation

To qualitatively evaluate how the causal intervention mechanism suppresses spurious correlations and enhances the interpretability of the model, this section provides a centralized analysis by integrating t-SNE visualizations and multimodal correlation heatmaps.

Feature Decoupling Analysis via t-SNE, Figure 8 illustrates the two-dimensional distribution of the 8 typical abnormal activity features in the OURS dataset. While the raw data and baseline models exhibit chaotic overlap (Figures 8 a-e) due to the interference of unobserved confounders, the features extracted by CIFAM (Figure 8 f) achieve a "clear decoupling" with compact clusters and significant inter-class physical distances. This transition qualitatively proves that by blocking confounding paths, CIFAM successfully separates core causal elements from unstable environmental background noise.

Cross-modal Deconfounding via Heatmaps, Figure 9 further characterizes the impact of the front-door adjustment strategy on feature correlations. In real ump station scenarios, mixed environmental noise artificially creates strong spurious correlations across different modalities (Figure 9a). After implementing CIFAM's causal intervention and contrastive constraints, the cross-modal correlation coefficients decrease significantly while intra-modal consistency is preserved (Figure 9b). This change indicates that the model effectively eliminates redundant background information and extracts independent feature representations that capture the stable causal essence of abnormal events, ensuring that the predictor follows physically realistic logic rather than statistical shortcuts.

## 5.8 Sample efficiency analysis

In the actual industrial scenario of pump station monitoring, high-quality anomaly-labeled data is often scarce, making a model's ability to leverage limited data a crucial criterion for its engineering value. In this section, we evaluate CIFAM against five representative baselines: the spatio-temporal models EvolveGCN and DySAT, the high-

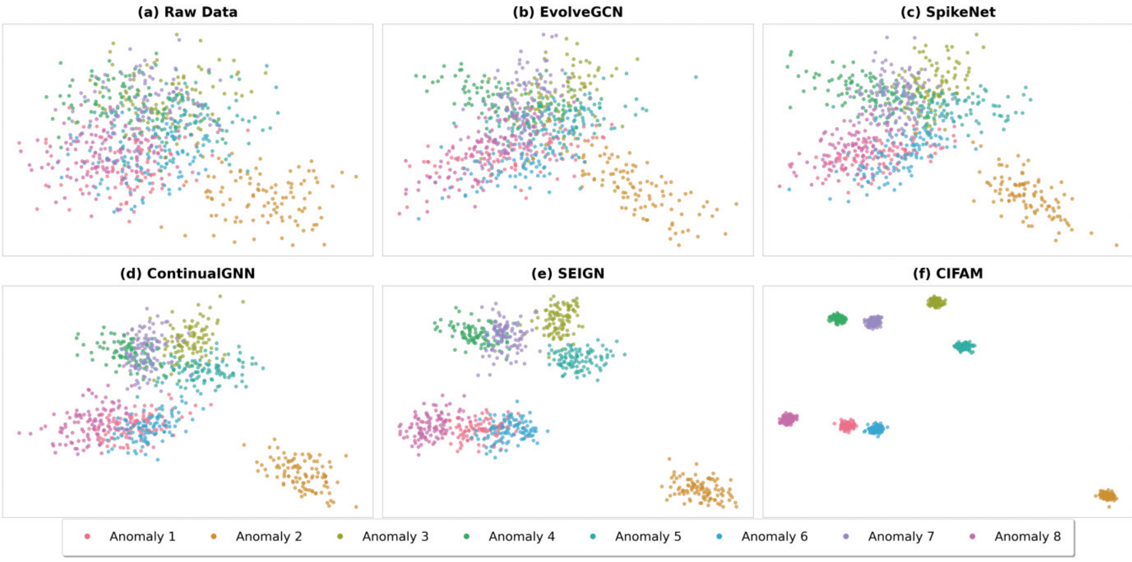


**Figure 8**. Comparison of t-SNE visualizations by different methods

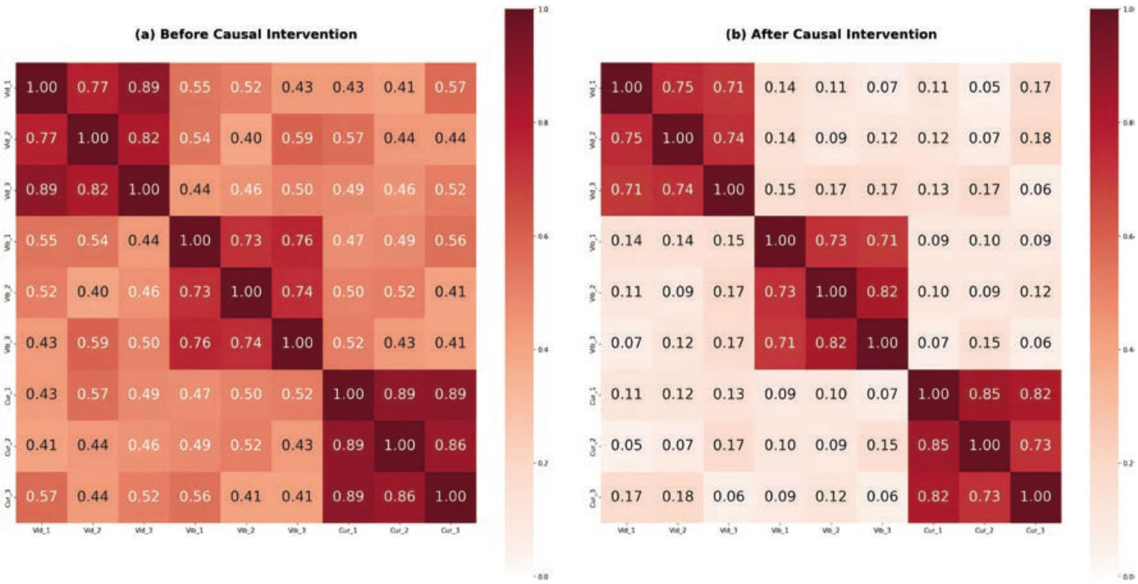


**Figure 9**. Heatmap of multimodal feature correlations before and after causal intervention

performance feature extraction model SpikeNet, and the recent causal-based frameworks CUBE and CIGNN.

As illustrated in Figure 10, all models exhibit a performance decline as the training data ratio decreases. However, traditional spatio-temporal models (EvolveGCN and DySAT) and the feature-driven SpikeNet suffer significantly from shortcut learning in low-data regimes, with their Micro-F1 scores dropping to approximately 66%-71% at the 10% training ratio. In contrast, causal-based methods demonstrate superior stability. CUBE, which utilizes counterfactual explanations, maintains an accuracy of approximately 74.5% under the 10% setting. Notably, CIFAM significantly outperforms all baselines, achieving a Micro-F1 of 78.5% even with minimal labeled data.

The superior sample efficiency of CIFAM is attributed to its front-door adjustment mechanism, which enables the model to identify the true causal mediator $\underline{Z}$ and block the interference of unobserved confounders prevalent in small datasets. While SpikeNet relies on massive samples to learn complex features, CIFAM's structural attribute graph and causal dilated convolution allow it to capture the underlying physical laws of pump station operations more directly. Consequently, CIFAM reaches performance saturation (around 88.4%) with only 40% of labeled data, demonstrating strong practical utility in data-scarce industrial environments.

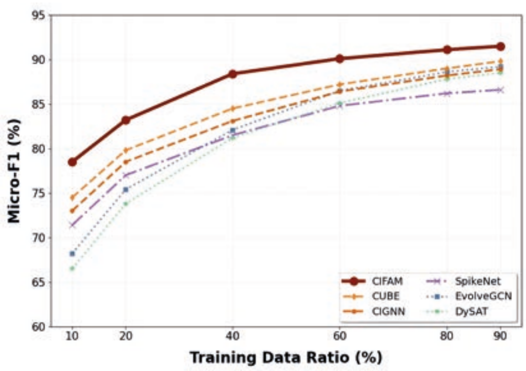


**Figure 10**. Sample efficiency analysis under different proportions of training data

## 5.9 A comparative analysis of methods for constructing structural attribute graphs

To verify the effectiveness of the multi-level structure attribute graph method in representing the dynamic dependency relationships of pump station operation activities, this section conducts a comparative analysis of the anomaly prediction performance of graph construction methods with different representation approaches. Four different graph construction strategies are selected: (1) Fully connected graph, which assumes that all pairs of modal features are related, resulting in a large number of redundancies in the graph; (2) K-nearest neighbor graph, which determines the adjacency relationship in the graph based on the similarity of dataset features, and only calculates the association between some feature pairs based on their similarity; (3) Static correlation graph, which sets the weights of the edges in the graph based on the Pearson correlation coefficient; (4) The multi-level structure graph proposed in this paper, which integrates video and signal features and realizes multi-level evolutionary structural associations through graph attention networks and dynamic clustering.

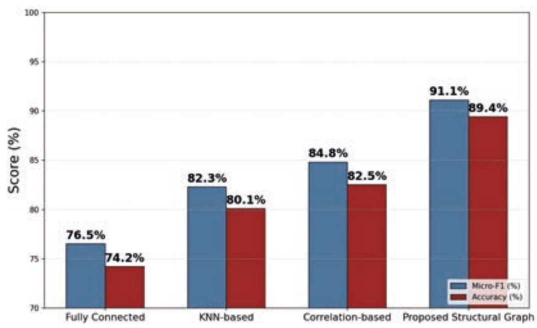


**Figure 11**. Performance comparison of different graph construction methods

As can be seen from Figure 11, when using a fully connected graph as the model input, the performance is poor, with a Micro-F1 of only 76.5%. This is because the false interaction between irrelevant nodes can introduce a large amount of noise. KNN graphs and correlation graphs can better remove redundancies, but since the graph structure is fixed or quasi-fixed, they cannot fully capture the various complex nonlinear dynamic evolution processes of the pumping station during operation. The graph construction strategy proposed in this paper is optimal in all indicators, with a Micro-F1 of up to 91.1%. This is because the multi-level clustering used to generate the structural graph more accurately restores the causal evolution path of the multimodal signals and removes some spatial redundancies through the structural graph, providing a better structural basis for subsequent causal intervention.

In the process of constructing a multi-layer structural attribute graph, the $K$ value determines the abstraction granularity and topological complexity of the graph structure. To evaluate the model's ability to capture the dynamic dependencies of multimodal data, we conducted comparative experiments on the OURS dataset by setting $K$ to 2, 4, 8, 16, 32, and 64 respectively. The results are shown in Figure 12.

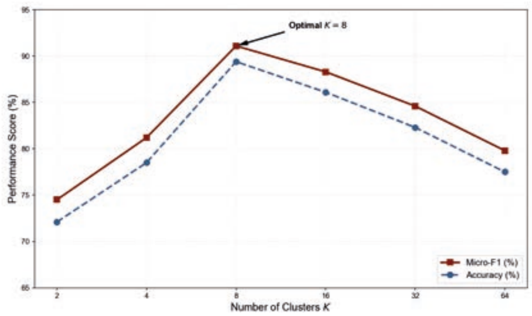


**Figure 12**. Sensitivity analysis of the number of clusters on model prediction performance

Experiments show that the model performance with respect to the $K$ value follows a pattern of ”first increase then decrease”. When $K$ is too small, the graph structure is too coarse to finely distinguish the fine-grained causal interaction relationships between molecular nodes, and it cannot effectively capture abnormal evolution paths. When $K = 8$, the model achieves the maximum values in Micro-F1 and Accuracy metrics, which is highly consistent with the eight types of abnormal activities and the physical state space set in this paper, indicating that this granularity strikes a balance between the representativeness of features and the simplicity of the structure. Beyond this point, although more clusters can increase the number of similar nodes within each cluster, it inevitably leads to overly fine clusters. Such overly fine clustering can cause fragmentation of the graph due to an excessive number of categories, introducing a large amount of topological noise and increasing the difficulty of graph attention calculation, without further improving the prediction accuracy. From the above experiments, it can be concluded that setting the clustering granularity reasonably is the basis for establishing a good structure-attribute graph and enabling causal intervention to take effect.

### 5.10 Efficiency and Complexity Analysis

To provide a quantitative evaluation of the model’s practical utility, this section analyzes the computational complexity and inference efficiency of CIFAM compared to mainstream baselines on the OURS dataset. The key metrics include parameter count (Params), floating-point operations (FLOPs), and average inference latency per sample.

**Table 2**. Comparison of Computational Complexity and Inference Efficiency

| Model | Params (M) | FLOPs (G) | Latency (ms) | Training Time (s) |
|---|---|---|---|---|
| EvolveGCN | 1.42 | 2.54 | 12.4 | 45.2 |
| DySAT | 1.85 | 3.12 | 15.6 | 58.7 |
| SpikeNet | 1.25 | 1.98 | 8.2 | 32.5 |
| CUBE | 2.1 | 4.05 | 22.1 | 75.4 |
| CIFAM | 1.38 | 2.21 | 9.5 | 38.6 |

As shown in the Table 2, CIFAM maintains a highly competitive computational profile. Despite the inclusion of front-door adjustment and causal dilated convolution, its inference latency (9.5 ms) is significantly lower than that of complex self-attention models like DySAT and counterfactual models like CUBE.

The efficiency of CIFAM stems from the design of the Causal Dilated Convolution module. Unlike traditional RNNs or deep CNNs that require massive layers to capture long-range dependencies, dilated convolution expands the receptive field exponentially with network depth. This allows the model to trace back to early operational parameters in time-lagged pump station signals without a linear increase in FLOPs. Furthermore, the Front-door Adjustment is implemented via a lightweight contrastive learning objective and AdaIN fusion, which primarily operates on graph-level representations rather than dense node-wise matrices, ensuring that the causal intervention does not affect real-time monitoring performance.

# Conclusion

This paper addresses the issue that anomaly prediction models in complex industrial scenarios such as pump stations are easily disturbed by unobservable confounding factors. It proposes and implements a multimodal causal intervention model based on the front-door adjustment strategy. By deeply integrating causal inference theory with deep learning techniques, this study achieves a transformation from statistical association to causal logical reasoning and constructs a multi-level structural attribute graph representation model. By integrating video dynamic features and sensor time series signals, it utilizes graph attention networks and clustering algorithms to capture the spatio-temporal evolution dependencies among multimodal data. A causal intervention and feature decoupling mechanism based on the front-door adjustment is proposed. In the case of unobservable confounding factors, random edge dropping is used to simulate the intervention process, and robust intermediate representations are extracted based on the front-door criterion. By combining causal dilated convolution and AdaIN techniques, it achieves a deep decoupling of core causal content and non-causal noise, blocking the propagation path of spurious correlations both theoretically and in model structure. Extensive experiments have verified the model's outstanding performance and generalization ability. On the self-built multimodal dataset of pump stations (OURS), the prediction accuracy of the CIFAM model reached 91.1%, significantly outperforming mainstream dynamic models and recent causal diagnosis frameworks such as CIGNN. Particularly in the sample efficiency experiment, the model maintained an accuracy of 78.5% even when only 10% of the labeled data was used. Additionally, through t-SNE visualization and correlation heatmaps, it qualitatively proves that the model successfully achieves redundancy removal between modalities, significantly enhancing the prediction stability under complex non-steady-state working conditions.

The CIFAM model proposed in this paper effectively eliminates spurious correlations in industrial scenarios through a front-door adjustment strategy, thereby enhancing the accuracy of anomaly prediction. In future work, we will further explore the universality of this model in more fields. Due to its core causal intervention mechanism and multi-modal fusion framework, which are highly portable, this method is not limited to pump station monitoring but can also be widely extended to the operation status monitoring of generator sets, the fine quality control of semiconductor production lines, and even the inspection and maintenance monitoring of long-distance infrastructure such as oil pipelines and high-voltage transmission lines using unmanned aerial vehicle-mounted multi-sensor equipment.

# Funding Statement

This work was supported by the National Natural Science Foundation, China (No. 61572035), the Natural Science Foundation of Anhui Province, China (No. 2308085US11), the Scientific Research Project of Anhui Provincial Education Department (2025AHGXZK10007), the Sub-project of National Key Science and Technology Program (BX2024C087), and Anhui Province Academic and Technical Leader Foundation (No. 2022D327), Postgraduate Innovation Fund(2024cx1013).

# Data Availability Statement

Data sets generated during the current study are available from the corresponding author on reasonable request.

# Conflicts of Interest

The authors declare that they have no competing interest.

# Author Contributions

Conceptualization, R. Z. and X. F.; Methodology, R. Z.; Software, R. Z.; Validation, R. Z., X. F. and K. L.; Formal Analysis, R. Z.; Investigation, R. Z.; Resources, X. F.; Data Curation, R. Z.; Writing – Original Draft Preparation, R. Z.; Writing – Review & Editing, R. Z.; Visualization, R. Z.; Supervision, K. L.; Project Administration, K. L.; Funding Acquisition, X. F.

## References


[1] S. B. Kaç, E. Sken, Sensor Fault Detection and RUL Estimation for Drinking Water Pumping Stations, IEEE Sensors Journal, 24, 2024, 15620-15629.

[2] C. E. Sunal, V. Dyo, V. Velisavljevic, Review of Machine Learning Based Fault Detection for Centrifugal Pump Induction Motors, IEEE Access, 10, 2022, 71344-71355.

[3] D. Chen, R. Liu, Q. Hu, et al., Interaction-Aware Graph Neural Networks for Fault Diagnosis of Complex Industrial Processes, IEEE Transactions on Neural Networks and Learning Systems, 34, 2023, 6015-6028.

[4] R. Liu, Q. Zhang, D. Lin, et al., Causal intervention graph neural network for fault diagnosis of complex industrial processes, Reliability Engineering & System Safety, 251, 2024, 110328.

[5] C. Gong, C. Zhang, D. Yao, Causal Discovery from Temporal Data: An Overview and New Perspectives, ACM Computing Surveys, 57, 2025, 4.

[6] J. J. Q. Yu, Graph Construction for Traffic Prediction: A Data-Driven Approach, IEEE Transactions on Intelligent Transportation Systems, 23, 2022, 15015-15027.

[7] H. Peng, H. Wang, B. Du, et al., Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting, Information Sciences, 521, 2020, 277-290.

[8] P. Mukherjee, T. C. Shen, J. Liu, et al., Confounding factors need to be accounted for in assessing bias by machine learning algorithms, Nature Medicine, 28, 2022, 1159-1160.

[9] R. Geirhos, J.-H. Jacobsen, C. Michaelis, et al., Shortcut learning in deep neural networks, Nature Machine Intelligence, 2, 2020, 665-673.

[10] L. Yue, Q. Liu, Y. Liu, et al., Learning from shortcut: a shortcut-guided approach for explainable graph learning, Frontiers of Computer Science, 19, 2025, 198338.

[11] H. Liu, H. Zhang, P. Li, Denoising Implicit Feedback for Graph Collaborative Filtering via Causal Intervention, IEEE Transactions on Big Data, 11, 2025, 696-709.

[12] Z. Tian, Z. Zhou, X. Hua, et al., CATER: Causal Adaptive Tracking with Environmental Reasoning, Signal Processing, 238, 2026, 110169.

[13] X. Deng, X. Tian, S. Chen, et al., Nonlinear Process Fault Diagnosis Based on Serial Principal Component Analysis, IEEE Transactions on Neural Networks and Learning Systems, 29, 2018, 560-572.

[14] M. Sugiyama, Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis, J. Mach. Learn. Res., 8, 2007, 1027-1061.

[15] O. Costilla-Reyes, P. Scully, K. B. Ozanyan, Deep Neural Networks for Learning Spatio-Temporal Features From Tomography Sensors, IEEE Transactions on Industrial Electronics, 65, 2018, 645-653.

[16] K. Zhong, M. Han, T. Qiu, et al., Fault Diagnosis of Complex Processes Using Sparse Kernel Local Fisher Discriminant Analysis, IEEE Transactions on Neural Networks and Learning Systems, 31, 2020, 1581-1591.

[17] D. Yang, Y. Pang, B. Zhou, et al., Fault Diagnosis for Energy Internet Using Correlation Processing-Based Convolutional Neural Networks, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49, 2019, 1739-1748.

[18] Z. C. Lipton, A Critical Review of Recurrent Neural Networks for Sequence Learning, arXiv, abs/1506.00019, 2015.

[19] E. L. Manibardo, I. Laña, J. D. Ser, Deep Learning for Road Traffic Forecasting: Does it Make a Difference?, IEEE Transactions on Intelligent Transportation Systems, 23, 2022, 6164-6188.

[20] J. Ye, J. Zhao, K. Ye, et al., How to Build a Graph-Based Deep Learning Architecture in Traffic Domain: A Survey, IEEE Transactions on Intelligent Transportation Systems, 23, 2020, 3904-3924.

[21] T. N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.

[22] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering, In: Neural Information Processing Systems, 2016.

[23] Z. Liu, J. Zhou, Graph Attention Networks, In: Introduction to Graph Neural Networks, Springer International Publishing, Cham, 2020, 39-41.

[24] L. Zhao, Y. Song, C. Zhang, et al., T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction, IEEE Transactions on Intelligent Transportation Systems, 21, 2020, 3848-3858.

[25] M. Wyrembek, G. Baryannis, A. Brintrup, Causal Machine Learning for Supply Chain Risk Prediction and Intervention Planning, International Journal of Production Research, 63, 2025, 5629-5648.

[26] Y. Niu, K. Tang, H. Zhang, et al., Counterfactual VQA: A Cause-Effect Look at Language Bias, In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 12695-12705.

[27] R. Liu, H. Liu, G. Li, et al., Contextual Debiasing for Visual Recognition with Causal Mechanisms, In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[28] Y. Liu, G. Li, L. Lin, Cross-Modal Causal Relational Reasoning for Event-Level Visual Question Answering, IEEE Trans Pattern Anal Mach Intell, 45, 2023, 11624-11641.

[29] W. Li, Z. Li, X. Yang, et al., Causal-ViT: Robust Vision Transformer by causal intervention, Engineering Applications of Artificial Intelligence, 126, 2023, 107123.

[30] D. Yu, Q. Li, X. Wang, et al., Deconfounded recommendation via causal intervention, Neurocomputing, 529, 2023, 128-139.

[31] X. Huang, S. Belongie, Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization, In: 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[32] T. Tanabe, J. Ye, T. Suzuyama, Potential for Improving the Local Realization of Coordinated Universal Time with a Convolutional Neural Network, Review of Scientific Instruments, 90, 2019, 125001.

[33] R. Xu, Z. Song, J. Wu, Change-Point Detection with Deep Learning: A Review, Frontiers of Engineering Management, 12, 2025, 154-176.

[34] Q. Li, D. Liu, J. Kong, Temporal Action Localization with Cross Layer Task Decoupling and Refinement, in: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI Press, 2025.

[35] K. P. Sinaga, M.-S. Yang, Unsupervised K-Means Clustering Algorithm, IEEE Access, 8, 2020, 80716-80727.

[36] Z. Liu, D. Yang, Y. Wang, et al., EGNN: Graph structure learning based on evolutionary computation helps more in graph neural networks, Applied Soft Computing, 135, 2023, 110040.

[37] Y. Tian, Y. Qi, F. Guo, FreeDyG: Frequency Enhanced Continuous-Time Dynamic Graph Model for Link Prediction, B. Kim, Y. Yue, S. Chaudhuri, et al., In: International Conference on Representation Learning, 2024.

[38] A. Sankar, Y. Wu, L. Gou, DySAT: Deep Neural Representation Learning on Dynamic Graphs via Self-Attention Networks, in: Proceedings of the 13th International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, NY, USA, 2020, pp. 519-527.

[39] A. Pareja, G. Domeniconi, J. Chen, T. Ma, T. Suzumura, H. Kanezashi, T. Kaler, T. Schardl, C. Leiserson, EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs, Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, 5363-5370.

[40] H. Lu, L. Wang, X. Ma, A Survey of Graph Neural Networks and Their Industrial Applications, Neurocomputing, 614, 2025, 128761.

[41] T. Yi, S. Zhang, Z. Bu, Link Prediction Based on Higher-Order Structure Extraction and Autoencoder Learning in Directed Networks, Knowledge-Based Systems, 241, 2022, 108241.

[42] J. Chen, Y. Yuan, X. Luo, SDGNN: Symmetry-Preserving Dual-Stream Graph Neural Networks, IEEE/CAA Journal of Automatica Sinica, 11, 2024, 1717-1719.

[43] F. M. Bianchi, D. Grattarola, L. Livi, Graph Neural Networks With Convolutional ARMA Filters, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44, 2022, 3496-3507.

[44] X. Shao, H. Wang, X. Chen, CUBE: Causal Intervention-Based Counterfactual Explanation for Prediction Models, IEEE Transactions on Knowledge and Data Engineering, 36, 2024, 2416-2429.

**Ruoyuan Zhang** received the M.S. degree from Computer Technology from North Minzu University in 2021,he is currently pursuing the Ph.D. degree in Information Security Engineering at the School of Mathematics and Big Data at Anhui University of Science and Technology.His research interests include deep learning, artificial intelligence, and anomaly detection.
https://orcid.org/0000-0002-2879-7539

**Xianwen Fang** received the M.S. degree from Shandong University of Science and Technology, China, in 2004, and the Ph.D. degree from the Key Laboratory of Service Computing, Tongji University, China, in 2011. He is currently working as a Professor with the Department of Computer Science and Engineering, School of Mathematics and Big Data, Anhui University of Science and Technology, China. His research interests include petri net theory, trustworthy software, and web services.
https://orcid.org/0000-0001-8531-7215

**Lu Ke** received the B.S. degree from Anhui University of Science and Technology, China, in 2017, and the Ph.D. degree from the same university in 2022. He has been working at the School of Mathematics and Big Data, Anhui University of Science and Technology since graduation. His research mainly focuses on Petri nets, online monitoring of business processes, and artificial intelligence.
https://orcid.org/0000-0002-0792-2628