

Operating performance assessment for complex industrial processes based on multi-block sparse spatiotemporal graph transformer

Ruicong Zhang, Fei Chu ^(✉), Shuyan Xu, and Fuli Wang

Abstract Energy and material transfer and physicochemical reactions in complex industrial processes exhibit spatiotemporal propagation characteristics, and spatiotemporal relationships among variables can reveal operating patterns and improve operating performance assessment. However, numerous redundant spatiotemporal relationships interfere with operating performance representation, reducing model interpretability and cross-distribution performance stability. To address this issue, this paper proposes an operating performance assessment method for complex industrial processes based on multi-block sparse spatiotemporal graph transformer. Specifically, a perturbation response mechanism based on a conditional variational autoencoder is designed to measure the sensitivity of edge relationships to local representation changes and select perturbation-responsive edges that carry critical operating performance discriminative information. Meanwhile, a masked decay matrix is designed to impose temporal decay constraints on spatiotemporal information propagation, reflecting the continuous evolution of industrial process variables. Finally, quantile thresholds are constructed based on the sparse dynamic graph obtained from the optimal operating performance level, and key non-optimal variables are traced by analyzing edge-intensity variations. Experimental results on a coal preparation process demonstrate that the proposed method can effectively prune weakly responsive redundant relationships, achieving a pruning rate of 93.42%, while maintaining high operating performance assessment accuracy and performance stability under cross-distribution scenarios.

Keywords process operating performance assessment, graph transformers, sparse graph, pruning strategy, coal preparation process.

1 Introduction

Complex industrial processes extract valuable target products from chemically stable raw materials in the natural environment through multiple forms of energy–material transfer and physicochemical reactions. These processes typically operate in complex open environments and consist of numerous interconnected production units, involving a large number of variables with intricately coupled spatiotemporal relationships. Any abnormality occurring in a single unit may propagate throughout the entire production line, thereby degrading overall operational efficiency and product quality. Consequently, timely and accurate process monitoring is crucial to ensuring product quality and production efficiency [1].

However, modern enterprises or plants usually pursue higher economic benefits. Therefore, process monitoring alone is insufficient to meet their demand for economic performance. Process operating performance assessment (POPA) enables more refined analysis by focusing on the degree to which a process deviates from the economically optimal operating performance level within the normal operating range, thereby ensuring that complex industrial processes operate as close as possible to the optimal operating performance level and improving overall economic benefits [2]. Over the past decade, data-driven POPA methods have become a mainstream research direction [3]. For different process characteristics, researchers have employed multivariate statistical techniques to jointly analyze multiple variables and their interactions, and to construct monitoring statistics in low-dimensional spaces [4]. However, with the increasing scale and complexity of modern industrial processes, multivariate

• Ruicong Zhang and Shuyan Xu are with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China. E-mail: tb23060022a41@cumt.edu.cn, ts25060243p31@cumt.edu.cn

• Fei Chu is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China; Underground Space Intelligent Control Engineering Research Center of the Ministry of Education, Xuzhou 221116, China. E-mail: chufei@cumt.edu.cn

• Fuli Wang is with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. E-mail: wangfuli@mail.neu.edu.cn

* To whom correspondence should be addressed.



清华大学出版社
Tsinghua University Press

SciOpen

statistical methods suffer from inherent limitations in data adaptability and feature extraction capability, which restrict their applicability in complex production environments.

Deep learning-based POPA models [5] characterize the complex relationships between process variables and operating performance levels through multilayer nonlinear mappings, and learn feature representations related to POPA from process data. To address challenges in process data, such as zero-shot scenarios [6], insufficient labeled samples [7], and image distortion [8], researchers have employed deep learning models including stacked denoising autoencoders, generative adversarial networks, and convolutional neural networks to achieve robust feature learning and efficient representation of complex process data. However, the above methods are primarily confined to the scope of static feature learning. It should be noted that practical industrial processes generally exhibit dynamic characteristics, and the resulting process data show complex temporal correlations, where temporal features are a core attribute for characterizing the evolution of system states [9]. Accordingly, Xu et al. [10] utilized long short-term memory (LSTM) networks to extract effective features from time-series data of the blast furnace ironmaking process, thereby accurately identifying the work conditions of the blast furnace. Chu et al. [11], considering the dynamic characteristics of industrial processes, proposed an LSTM network based on supervised slow feature analysis for operating performance assessment in a dense medium coal preparation process. However, in dynamic processes, process variables are not only highly correlated over time but also interrelated in space [12].

Graph structures can represent the association patterns among variables in the form of nodes and edges [13]. Learning and exploiting such structural features in non-Euclidean space helps capture the complex spatial dependencies among variables, thereby improving the accuracy and robustness of POPA. One line of research adopts static graphs to model variable relationships by encoding the relatively fixed topological associations among process variables into graph structures, so as to alleviate the limitation of traditional temporal models in characterizing spatial coupling among variables [14] [15]. However, the relationships among variables usually change dynamically with operating condition shifts, operation adjustments, and other factors. Therefore, to describe the time-varying spatial dependency relationships among variables more precisely, Duan et al. [16] constructed dynamic topological graphs using a continuous sliding window and mutual information and applied them to early fault warning in chemical processes. Zhang et al. [17] proposed a corre-

lation graph structure learning module based on the spatial self-attention mechanism to adaptively learn the correlation graph.

However, the aforementioned dynamic graphs mainly characterize the associations among variables, and the learned observational relationships may not necessarily benefit downstream tasks. Recently, task-oriented adaptive graph construction methods have attempted to directly optimize graph structures through feedback from downstream tasks. For example, Dong et al. [18] employed graph reinforcement learning to adaptively determine edge connections and edge weights among nodes, and optimized the graph construction strategy based on feedback from fault diagnosis performance. Li et al. [19] further combined dual-level graph representations with graph reinforcement learning from the perspective of graph reconstruction, and optimized graph structures through reward feedback. These studies indicate that graph construction should not only characterize potential relationships among objects but also consider the practical representation value of different structural relationships for downstream tasks.

For complex industrial process operating performance assessment, not all variable relationships have the same value for operating performance representation [20]. Dynamic graphs constructed from observational data may contain numerous redundant relationships, causing relationships with limited contributions to operating performance discrimination to participate in node neighborhood aggregation, thereby affecting model interpretability and cross-distribution performance stability. In contrast, edge relationships carrying critical operating performance discriminative information can respond more sensitively to local representation changes and influence information propagation among nodes through edge-weight adjustment, enabling effective transmission of critical operating performance discriminative information. Therefore, it is necessary to identify effective relationships that contribute to operating performance discrimination from numerous candidate relationships and construct a more effective neighborhood information propagation structure.

In summary, for complex industrial processes with highly coupled and varying spatiotemporal relationships, it is necessary to select valuable information for operating performance representation to reduce the interference of redundant spatiotemporal relationships on POPA and non-optimal variable tracing. Therefore, the following challenges need to be addressed:

- 1) At the spatial relationship modeling level, message passing mechanisms assume that the observed graph is reliable. However, not all edge relationships carry critical operating

performance discriminative information, and redundant relationships introduce low-value information propagation while increasing graph structural complexity.

2) At the temporal relationship modeling level, variable states in continuous operating processes are affected by the cumulative effects of historical moments [21]. If temporal direction and temporal decay constraints are ignored and historical information is modeled equivalently, redundant spatiotemporal dependencies that do not conform to the process evolution law may be introduced.

3) At the non-optimal variable tracing level, the strength of relationships among variables dynamically changes with process evolution. However, abundant redundant relationships interfere with the construction of baseline thresholds, making it difficult to identify key non-optimal variables and increasing the interpretability complexity of non-optimal variable tracing.

To address the challenges caused by redundant spatiotemporal relationships in POPA, this paper proposes an operating performance assessment method for complex industrial processes based on multi-block sparse spatiotemporal graph transformer (MSSGformer). Specifically, during dynamic graph construction, a conditional variational autoencoder (CVAE)-based perturbation response mechanism (CVAE-PRM) is designed to measure the sensitivity of edge relationships to local representation changes and select perturbation-responsive edges carrying critical operating performance discriminative information, thereby constructing a sparse dynamic graph. Subsequently, the graph convolutional neural network (GCN) is employed to capture the spatial dependency features of variables. Meanwhile, considering the temporal accumulation effect of variables, a masked decay matrix (MDM) is designed to propagate spatiotemporal features under temporal decay constraints, which is consistent with the objective law of continuous evolution of industrial process variables. For non-optimal operating performance, key non-optimal variables are traced by examining the intensity changes of edge relationships. Finally, the effectiveness of the proposed method is verified on a coal preparation process. The main contributions of this work are summarized as follows:

1) A CVAE-based perturbation response mechanism is designed to measure the sensitivity of dynamic edge relationships, select perturbation-responsive edges carrying critical operating performance discriminative information, and achieve POPA-oriented dynamic graph sparsification.

2) A masked decay matrix consistent with the physical evolution law of industrial processes is constructed to transmit spatiotemporal features with temporal decay constraints.

3) Adaptive decision thresholds are constructed from the sparse dynamic graph under the optimal operating performance level, and key non-optimal variables are traced by analyzing edge relationship intensity variations.

The rest of this article is organized as follows. Section 2 briefly introduces the CVAE-PRM, the graph transformers, and the MDM. Section 3 presents the proposed POPA framework. Section 4 validates the effectiveness of the proposed method on a coal preparation process. Finally, Section 5 provides the conclusion.

2 Preliminaries

2.1 CVAE-based Perturbation Response Mechanism

In complex industrial processes, not all edge relationships have the same value for operating performance discrimination. If a local process representation undergoes a reasonable variation and an edge relationship produces a significant response, it indicates that the edge can map local representation changes into adjustments of inter-variable relationship strength and regulate neighborhood information propagation through edge-weight variations during graph message passing. Therefore, such edges can more effectively reflect operating performance variation characteristics and carry critical operating performance discriminative information. In contrast, weakly responsive relationships are insensitive to local representation changes, resulting in limited adjustments of relationship strength and lower contributions to operating performance discrimination. Moreover, they may introduce redundant neighborhood information and interfere with node feature aggregation. Therefore, this paper measures the response degree of edge relationships to local representation changes, selects perturbation-responsive edges, and treats weakly responsive relationships as pruning candidates.

Since local generation relationships in complex industrial processes are difficult to directly obtain, CVAE is employed to approximate the conditional generative representations of variables within the model. For variable X_i , the CVAE infers its latent variable Z_i through the encoder and characterizes the conditional generative representation of X_i under the joint effect of the conditional input Pa_i and the latent variable Z_i through the decoder. This process can be expressed as:

$$p_{\theta}(X_i | Pa_i) = \int p_{\theta}(X_i | Z_i, Pa_i) p(Z_i) dZ_i \quad (1)$$

where Z_i denotes the latent variable, and $p_{\theta}(X_i | Z_i, Pa_i)$ denotes the conditional generation distribution characterized by the decoder.

To construct controlled variations of local generation representations, a perturbation is introduced into the latent space

of the CVAE:

$$Z_i^\pi = Z_i + \Delta Z_i \quad (2)$$

where ΔZ_i denotes the perturbation term. The model conditional generation process after perturbation can be expressed as:

$$p_\theta^\pi(X_i | Pa_i) = \int p_\theta(X_i | Z_i^\pi, Pa_i) p_\pi(Z_i^\pi) dZ_i^\pi \quad (3)$$

The latent-space perturbation induces changes in the local generation representation. While keeping the conditional input unchanged, the perturbed dynamic graph is reconstructed, and the variation in edge relationship strength before and after perturbation is compared to measure the response degree of different edge relationships to local representation changes. For edge e_{ij} , its perturbation response strength is defined as:

$$\Delta a_{ij} = |a_{ij}^\pi - a_{ij}| \quad (4)$$

where a_{ij} and a_{ij}^π represent the edge relationship strengths before and after perturbation, respectively, and Δa_{ij} represents the perturbation response strength of edge e_{ij} to local representation changes. A larger Δa_{ij} indicates that the edge can more sensitively respond to local representation variations and adjust information aggregation among nodes through relationship strength changes.

2.2 Graph transformers

The synergy between Transformers and graph learning has demonstrated strong performance and broad applicability in various graph-related tasks [22]. A graph transformer can simultaneously integrate the spatial feature extraction capability of graph neural networks and the temporal feature modeling capability of Transformers, thereby enabling a unified representation of spatiotemporal dependencies. As a variant of graph neural networks [23], graph convolutional networks update the state of each node by aggregating neighborhood information and updating target node representations [24]. For a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ denotes the set of N nodes, $\mathcal{E}$ denotes the set of edges, and $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix. If $A_{ij} \neq 0$, then $e_{ij} \in \mathcal{E}$, indicating a directed connection from v_i to v_j . The message passing scheme can be defined as:

$$H^{l+1} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^l W^l \right) \quad (5)$$

where $\tilde{A} = A + I_N$ represents the adjacency matrix with self-loops added, $\tilde{D}$ is the corresponding degree matrix, H^l and H^{l+1} denote the node feature representations at the l -th and $(l+1)$ -th layers, respectively, W^l is the trainable weight matrix of the l -th layer, and $\sigma(\cdot)$ denotes the activation function.

The Transformer [25] is a model based on the self-attention mechanism, which can efficiently capture global dependencies and exhibits strong parallel computation capability. The multi-head attention mechanism of the Transformer enables parallel attention to information from different subspaces and can be expressed as:

$$\text{head}_i = \text{softmax} \left(\frac{Q_i K_i^\top}{\sqrt{d_k}} \right) V_i \quad (6)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W_o \quad (7)$$

where Q , K , and V denote the query, key, and value matrices, respectively. head_i denotes the i -th self-attention head. d_k is the scaling factor, W_o is the learnable projection matrix after concatenation, $\text{Concat}(\cdot)$ denotes the concatenation operation, and $\text{softmax}(\cdot)$ denotes the normalized exponential function.

2.3 Masked Decay Matrix

In multivariate time series, it is generally assumed that temporal correlations among observations tend to decay as the time distance between their timestamps increases [21]. From the spatial perspective, a node's current state is typically more strongly correlated with its states at adjacent time steps. To characterize such time-decaying dependencies, the decay matrix can be used to constrain the temporal transmission strength of information. The decay matrix can be defined as $C = \{c_{tr}\}_{t,r=1}^T$, where each element $c_{tr} = \delta^{|t-r|}$. Here, δ denotes the decay rate, which is used to control the attenuation speed of information correlation as the temporal interval increases.

Meanwhile, considering that time-series data contain rich dynamic evolution information in the temporal dimension, temporal priority is a basic assumption [26]. Therefore, historical state information usually has a guiding effect on the current state, which is reflected in both temporal and spatial dimensions. In the temporal dimension, the representation of a time node is associated with historical temporal information. In the spatial dimension, the representation of a variable node is related to both the interaction relationships at the current time step and its historical spatial information. Therefore, to characterize such spatiotemporal information propagation that evolves over time, this paper designs a masked decay matrix. This matrix not only considers the influence of temporal distance on information transmission strength, but also explicitly determines the direction of information transmission through

the masking mechanism. It can be expressed as:

$$M = \{m_{tr}\}_{t,r=1}^T, \quad m_{tr} = \begin{cases} \delta^{t-r}, & \text{if } 0 \leq t-r \leq d_{\max}, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

where $d_{\max}$ denotes the maximum information propagation distance. This matrix ensures a reasonable strength of information transmission and effectively prevents backward leakage of future information into current or historical states, thereby ensuring that information transmission follows the chronological order, as illustrated in Fig. 1.

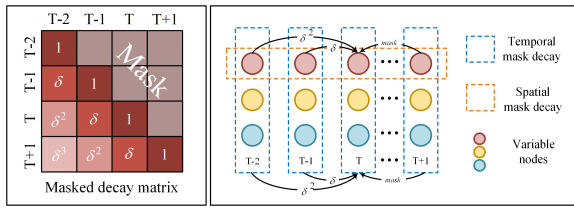


Fig. 1 Masked decay matrix.

3 Methodology

This section introduces the proposed MSSGformer-based operating performance assessment framework. First, the MSSGformer model is introduced. Then, non-optimal variables are traced for non-optimal operating performance. The overall framework is illustrated in Fig. 2.

In this work, we address the following tasks. Given the collected process data $\mathcal{X} = \{x_1, x_2, \dots, x_N\} \in \mathbb{R}^{T \times N \times d}$, where T denotes the number of time steps, N is the number of variables, and d represents the dimensionality of each variable. Since global joint modeling tends to introduce redundant cross-subprocess spatiotemporal dependencies, increase modeling complexity, and weaken local feature representation, the process is decomposed into L sub-blocks $\mathcal{B} = \{b_1, b_2, \dots, b_L\}$ according to a predefined partitioning criterion. The samples of the l -th sub-block can be expressed as $b_l = \{x_l^1, x_l^2, \dots, x_l^{n_l}\} \in \mathbb{R}^{T \times n_l \times d}$, where n_l denotes the number of variables in sub-block l . For each sub-block, a dynamic spatiotemporal feature representation model is constructed. Finally, the representations of all sub-blocks are fused to perform global POPA. Our objective is to learn a mapping function f to achieve POPA $\hat{y} = f(\mathcal{X})$, such that the evaluation risk $\mathbb{E}_{(\mathcal{X}, y)}[\mathcal{L}(f(\mathcal{X}), y)]$ is minimized, where y and $\hat{y}$ denote the actual and predicted process operating performance levels, respectively, and $\mathcal{L}(\cdot)$ represents the assessment loss function.

3.1 MSSGformer Model

The proposed MSSGformer model consists of three parts: the sub-block spatial relationship representation module (SRRM), the sub-block temporal relationship representation module (TRRM), and the process operating performance assessment module (POPAM). The SRRM is used to construct sparse dynamic graphs within each sub-block and select perturbation-responsive edges through the CVAE-PRM. The TRRM is used to characterize the dynamic evolution features in sub-block time series. The POPAM further integrates the spatial relationship features and temporal relationship features of each sub-block to achieve global POPA.

SRRM: In complex industrial processes, inter-variable dependency relationships dynamically adjust with operating conditions and process variations. To characterize such dependency relationships, MSSGformer first constructs an attention-based dynamic graph structure within each sub-block. For the data $b_{l,t} \in \mathbb{R}^{n_l \times d}$ of sub-block l at any time t , the correlation adjacency matrix between variables is calculated through the self-attention mechanism:

$$A_{l,t} = \text{softmax} \left(\frac{Q_{l,t} K_{l,t}^\top}{\sqrt{d}} \right) \quad (9)$$

where $Q_{l,t} = b_{l,t} W_{l,Q}$, $K_{l,t} = b_{l,t} W_{l,K}$, and $A_{l,t} \in \mathbb{R}^{n_l \times n_l}$. $A_{l,t}$ reflects the dynamic relationship strength among variables within the current sub-block.

Although complete mechanism modeling of complex industrial processes is highly difficult, their basic process flows, material transmission directions, and partial physical constraints can usually be provided by process knowledge. Therefore, an expert masking matrix (EMM) $M_E \in \{0, 1\}^{n_l \times n_l}$ is defined to constrain the information transmission direction within each sub-block, resulting in the expert-masked adjacency matrix $\tilde{A}_{l,t}$:

$$\tilde{A}_{l,t} = A_{l,t} \odot M_E \quad (10)$$

To avoid homogenization of conditional inputs for different nodes, node-specific conditional inputs are constructed based on $\tilde{A}_{l,t}$. For node $x_{l,t}^i$, its conditional input is obtained by weighting the neighbors related to this variable:

$$W_{l,t}^{ij} = \begin{cases} \frac{\tilde{A}_{l,t}(j, i)}{\sum_{k \neq i} \tilde{A}_{l,t}(k, i)}, & j \neq i, \\ 0, & j = i, \end{cases} \quad (11)$$

$$pa_{l,t}^i = \text{diag}(W_{l,t}^i) b_{l,t} \quad (12)$$

where $\text{diag}(W_{l,t}^i)$ denotes the diagonal matrix constructed by taking the elements of the vector $W_{l,t}^i$ as the diagonal entries.

On this basis, CVAE-PRM is designed to measure the sensitivity of dynamic edge relationships to local representation

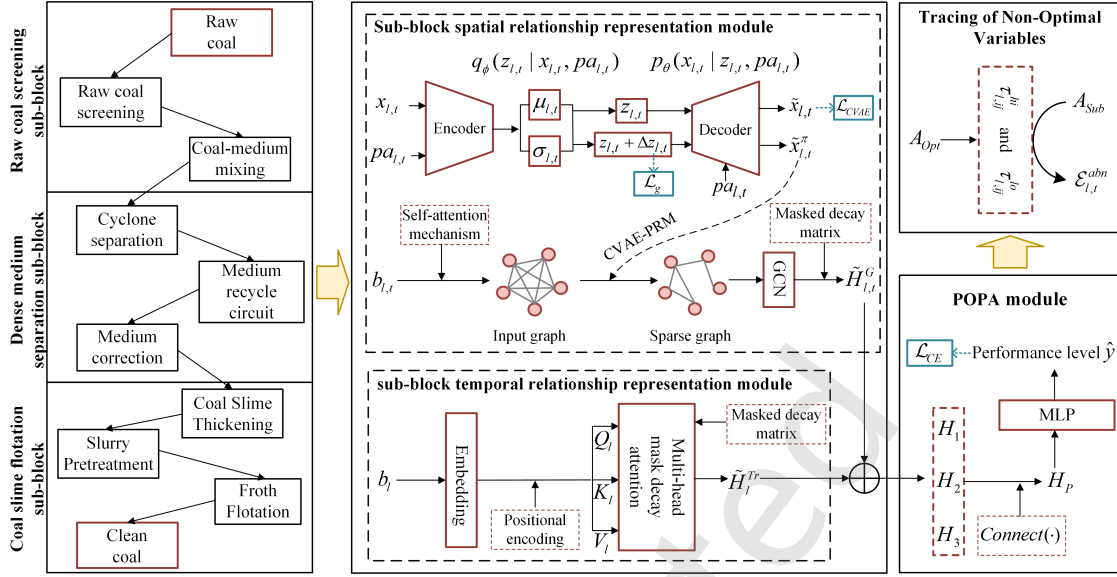


Fig. 2 MSSGformer-based POPA framework.

changes and select perturbation-responsive edges carrying critical operating performance discriminative information. Since the true generation mechanism of complex industrial processes is difficult to obtain directly, this paper adopts a CVAE to approximate the generation representation of variables under conditional input constraints within the model. For variable $x_{l,t}^i$, the CVAE infers the latent variable $z_{l,t}^i$ through the encoder $q_\phi(z_{l,t}^i | x_{l,t}^i, pa_{l,t}^i)$, and characterizes the conditional generation distribution through the decoder $p_\theta(x_{l,t}^i | z_{l,t}^i, pa_{l,t}^i)$, where $pa_{l,t}^i$ denotes the conditional input representation of variable $x_{l,t}^i$, and $z_{l,t}^i$ denotes the latent variable. The training loss of the CVAE is:

$$\mathcal{L}_{\text{CVAE}} = -\mathbb{E}_{q_\phi} [\log p_\theta(x_{l,t}^i | z_{l,t}^i, pa_{l,t}^i)] + D_{\text{KL}}(q_\phi(z_{l,t}^i | x_{l,t}^i, pa_{l,t}^i) \| p(z_{l,t}^i)) \quad (13)$$

where the first term is the reconstruction error, which constrains the generated result to approach the original input, and the second term is the KL divergence, which constrains the posterior distribution to approach the prior distribution.

To analyze the response of dynamic edge relationships to local representation changes, a controlled perturbation is introduced into the latent space:

$$z_{l,t}^{i,\pi} = z_{l,t}^i + \Delta z_{l,t}^i \quad (14)$$

where $\Delta z_{l,t}^i$ is a learnable latent offset term, obtained by mapping the latent representation $z_{l,t}^i$ through $g_\psi(\cdot)$:

$$\Delta z_{l,t}^i = \frac{g_\psi(z_{l,t}^i)}{\|g_\psi(z_{l,t}^i)\|_2 + \varepsilon} \quad (15)$$

After perturbation, the conditional input $pa_{l,t}^i$ remains unchanged, while the latent variable $z_{l,t}^i$ is transformed into

$z_{l,t}^{i,\pi}$. Therefore, the corresponding local generation representation of the decoder changes. The perturbed reconstruction is expressed as:

$$\tilde{x}_{l,t}^{i,\pi} = f_\theta(z_{l,t}^{i,\pi}, pa_{l,t}^i) \quad (16)$$

Meanwhile, the original reconstruction under the unperturbed condition is expressed as:

$$\tilde{x}_{l,t}^i = f_\theta(z_{l,t}^i, pa_{l,t}^i) \quad (17)$$

where $f_\theta(\cdot)$ is the decoder mapping function, $\tilde{x}_{l,t}^{i,\pi}$ represents the reconstructed representation after perturbation, and $\tilde{x}_{l,t}^i$ represents the reconstructed representation without perturbation.

Furthermore, the perturbed reconstructed representations of all variables are combined to obtain the reconstructed representation of the sub-block $\tilde{b}_{l,t}^\pi$:

$$\tilde{b}_{l,t}^\pi = \text{Concat}(\tilde{x}_{l,t}^{1,\pi}, \tilde{x}_{l,t}^{2,\pi}, \dots, \tilde{x}_{l,t}^{n_l,\pi}) \quad (18)$$

Subsequently, the perturbed adjacency matrix is recalculated as $\tilde{A}_{l,t}^\pi$:

$$\tilde{A}_{l,t}^\pi = f_{\text{att}}(\tilde{b}_{l,t}^\pi) \odot M_E \quad (19)$$

where $f_{\text{att}}(\cdot)$ denotes the attention mapping operation. By comparing the changes in edge relationship strength before and after perturbation, the perturbation response matrix of edge relationships is defined as:

$$\Delta A_{l,t} = |\tilde{A}_{l,t}^\pi - \tilde{A}_{l,t}| \quad (20)$$

where $\Delta A_{l,t,ij}$ represents the perturbation response strength of edge $e_{l,t,ij}$ to local representation changes. A larger $\Delta A_{l,t,ij}$ indicates that the edge relationship can more effectively reflect local representation variations through ad-

justments in relationship strength and influence information aggregation among nodes through edge-weight changes. Therefore, such edges possess higher operating performance discriminative value.

To ensure that the perturbed reconstruction representation maintains the original data flow while triggering sufficient edge-relation responses, $g_\psi(\cdot)$ is optimized by the following loss function:

$$\mathcal{L}_g = \left\| \tilde{x}_{l,t}^{i,\pi} - \tilde{x}_{l,t}^i \right\|_2^2 - \left\| \tilde{A}_{l,t}^\pi - \tilde{A}_{l,t} \right\|_F^2 \quad (21)$$

where the first term constrains the perturbed reconstruction result from deviating too far from the original reconstruction representation, ensuring that the perturbation remains within a reasonable data representation space. The second term encourages dynamic graph edge relationships to produce observable responses under controlled perturbation.

After obtaining the perturbation response matrix $\Delta A_{l,t}$, perturbation-responsive edges are selected through the perturbation response threshold. The perturbation response threshold is adaptively determined according to the statistical characteristics of ΔA_l within the current input time window. Specifically, the mean μ_l and standard deviation σ_l are calculated based on all valid nonzero elements in ΔA_l , and the perturbation response threshold is defined as:

$$\theta_l = \mu_l + \alpha \sigma_l \quad (22)$$

where α is the threshold control coefficient. Then, a selection mask is constructed according to the perturbation response strength:

$$C_{l,t,ij} = \begin{cases} 1, & \Delta A_{l,t,ij} > \theta_l, \\ 0, & \text{otherwise.} \end{cases} \quad (23)$$

The mask is used to retain perturbation-responsive edges with significant responses to local representation changes and remove weakly responsive redundant relationships, thereby obtaining the sparsified dynamic adjacency matrix $\tilde{A}_{l,t,ij}^c$:

$$\tilde{A}_{l,t,ij}^c = C_{l,t,ij} \odot \tilde{A}_{l,t,ij} \quad (24)$$

In addition, to reduce directional redundancy caused by simultaneous activation of bidirectional relationships, when both (i, j) and (j, i) pass the response-based selection, only the direction with a larger perturbation response magnitude is retained. The final sparse dynamic adjacency matrix $A_{l,t}^c$ is obtained as:

$$A_{l,t,ij}^c = \begin{cases} \tilde{A}_{l,t,ij}^c, & \Delta A_{l,t,ij} \geq \Delta A_{l,t,ji}, \\ 0, & \Delta A_{l,t,ij} < \Delta A_{l,t,ji}. \end{cases} \quad (25)$$

After obtaining the sparse dynamic adjacency matrix $A_{l,t}^c$, a GCN is employed to learn the spatial dependency features within the sub-block, which can be expressed as:

$$H_{l,t}^G = \sigma(A_{l,t}^c b_{l,t} W_{l,g}) \quad (26)$$

where $W_{l,g}$ is a learnable weight matrix, and $H_{l,t}^G$ represents the spatial dependency features of sub-block l at time step t .

Finally, to fuse spatial information from historical time steps, the masked decay matrix M is introduced, through which node representations are further aggregated in a weighted manner along the spatial dimension to obtain $\tilde{H}_{l,t}^G$:

$$\tilde{H}_{l,t}^G = \sum_{r=t_{\min}}^t m_{tr} H_{l,r}^G \quad (27)$$

where $\tilde{H}_{l,t}^G$ denotes the spatial relationship features obtained by the SRRM. $t_{\min} = \max(1, t - d_{\max})$. m_{tr} denotes the element at the t -th row and r -th column of the masked decay matrix M , representing the decay strength of the influence of historical moments on the current moment.

TRRM: Complex industrial processes exhibit continuously evolving dynamic characteristics, and the representation of a time node is related to historical temporal information. For any sub-block l , the temporal sequence input is denoted as $b_l \in \mathbb{R}^{T \times n \times d}$.

First, the original sequence is embedded into a d_{model} dimensional representation space through a linear projection and positional encoding, which can be expressed as:

$$H_{l,E} = b_l W_{l,e} + PE(t) \quad (28)$$

where $W_{l,e}$ is a learnable embedding matrix, and $PE(t)$ denotes the positional encoding.

In the encoder layer, multi-head self-attention is applied to $H_{l,E}$. The attention weight matrix can be expressed as:

$$S_l = \text{softmax} \left(\frac{Q_l K_l^\top}{\sqrt{d_{\text{model}}}} \right) \quad (29)$$

where $Q_l = H_{l,E} W_l^Q$ and $K_l = H_{l,E} W_l^K$. S_l represents the attention weight matrix among time steps, and its element $S_{l,tr}$ denotes the basic information aggregation weight from historical time step r to the current time step t .

Considering that the evolution of industrial process variables is subject to an obvious temporal precedence constraint, and that the influence of historical information on the current state usually decays gradually as the temporal distance increases, the masked decay matrix M is introduced to impose temporal constraints on the attention weights:

$$\tilde{S}_l = S_l \odot M \quad (30)$$

where $\tilde{S}_l$ represents the attention weights after incorporating temporal evolution constraints. This operation adjusts the propagation intensity of information with different temporal distances while preserving the data-driven temporal dependencies, enabling the model to focus more on historical information that is closely related to the current state and conforming to the temporal evolution characteristics of industrial

processes. Therefore, the attention output is obtained as:

$$H_l^{Tr} = \frac{\tilde{S}_{l,tr}}{\sum_r \tilde{S}_{l,tr}} V_l \quad (31)$$

By concatenating the outputs of all attention heads and applying linear projection, the output of multi-head mask attention can be obtained:

$$\tilde{H}_l^{Tr} = \text{Concat}(H_{l,1}^{Tr}, \dots, H_{l,h}^{Tr}) W_O \quad (32)$$

where $\tilde{H}_l^{Tr}$ denotes the temporal relationship features obtained by the TRRM. W_O is the output projection matrix, and h denotes the number of attention heads.

POPAM: The TRRM and SRRM extract complementary spatiotemporal features from the perspectives of variable relationships and temporal evolution, respectively. Based on this, the POPAM performs process-level feature fusion among all sub-block representations to achieve global POPA. For the l -th sub-block, its spatial relationship features and temporal relationship features are fused as:

$$H_l = \tilde{H}_l^{Tr} + \tilde{H}_{l,t}^G \quad (33)$$

Then, the fused features of all sub-blocks are concatenated along the variable dimension to obtain the global process representation H_P :

$$H_P = \text{Concat}(H_1, H_2, \dots, H_L) \quad (34)$$

Finally, the process-level features are fed into a multilayer perceptron (MLP) to implement POPA:

$$\hat{y} = f_{MLP}(H_P; \Theta) \quad (35)$$

where Θ denotes the network parameters, $\hat{y} \in \mathbb{R}^C$ represents the predicted probability distribution over operating performance, C denotes the number of classes, and $f_{MLP}(\cdot)$ denotes the multilayer perceptron mapping function.

Based on the prediction results, the cross-entropy loss function of POPA can be expressed as:

$$\mathcal{L}_{CE} = -\frac{1}{T} \sum_{i=1}^T \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}) \quad (36)$$

where $\hat{y}_{i,c}$ denotes the predicted probability of the c -th class for the i -th sample.

In summary, the total loss of MSSGformer is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_g \mathcal{L}_g + \lambda_{CVAE} \mathcal{L}_{CVAE} \quad (37)$$

where λ_g and λ_{CVAE} are the weighting coefficients of $\mathcal{L}_g$ and $\mathcal{L}_{CVAE}$, respectively. By minimizing $\mathcal{L}_{total}$, the parameters of the dynamic spatiotemporal feature learning module and the POPA model are jointly optimized, thereby realizing global POPA for complex industrial processes.

3.2 Tracing of Non-Optimal Variables

In complex industrial processes, deviations of operating performance from the optimal operating performance level are usually accompanied by changes in the strength of variable dependency relationships. Therefore, after model training, the key non-optimal variables can be traced by analyzing the variations of edge relationship intensities in the selected sparse dynamic graph, as shown in Fig. 3. This process not only identifies changes in the operating performance, but more importantly transforms operating performance deviation information into interpretable results at the variable level, providing operators with instructive analytical evidence to support engineering decisions such as abnormality localization and process adjustment.

Specifically, the sparse dynamic adjacency matrices obtained from the optimal operating performance level are first utilized to construct reference intervals of edge intensities. Let the dataset collected under optimal operating performance level be $\mathcal{X}_{opt}$. The set of sparse dynamic adjacency matrices of sub-block l under optimal operating performance level is defined as $A_l^{opt} = \{A_{l,t}^{c,opt} \mid t = 1, \dots, T_{opt}\}$, where T_{opt} denotes the total number of time steps under optimal operating performance level. For each edge (i, j) , the relationship strength sequence $\mathcal{E}_{l,ij}^{opt} = \{A_{l,t,ij}^{c,opt}\}$ is obtained, and its upper and lower quantile bounds are respectively defined as:

$$\tau_{l,ij}^{hi} = Q_{0.99}(\mathcal{E}_{l,ij}^{opt}), \quad \tau_{l,ij}^{lo} = Q_{0.01}(\mathcal{E}_{l,ij}^{opt}) \quad (38)$$

Accordingly, for the sparse dynamic adjacency matrix $A_{l,t}^{c,test}$ at any time instant t during the actual testing stage, its edge weight $A_{l,t,ij}^{c,test}$ is compared with the reference thresholds:

$$e_{l,t,ij}^{abn} = \begin{cases} 1, & A_{l,t,ij}^{c,test} > \tau_{l,ij}^{hi} \text{ or } A_{l,t,ij}^{c,test} < \tau_{l,ij}^{lo}, \\ 0, & \text{otherwise.} \end{cases} \quad (39)$$

where $e_{l,t,ij}^{abn} = 1$ indicates that the relationship intensity of edge (i, j) at time step t significantly deviates from the reference range established under optimal operating performance level. Based on this, the abnormal edge set at time step t can be defined as $\mathcal{E}_{l,t}^{abn} = \{(i, j) \mid e_{l,t,ij}^{abn} = 1\}$, which reveals the temporal evolution of abnormal edges and provides a basis for tracing key non-optimal variables.

4 Industrial Case Study

4.1 Coal Preparation Process and Dataset Description

Description of the coal preparation process: The coal preparation process is a typical complex industrial process system. Its operating environment is characterized by strong disturbances and high coupling, making the quality of clean coal susceptible to fluctuations and external interference. The

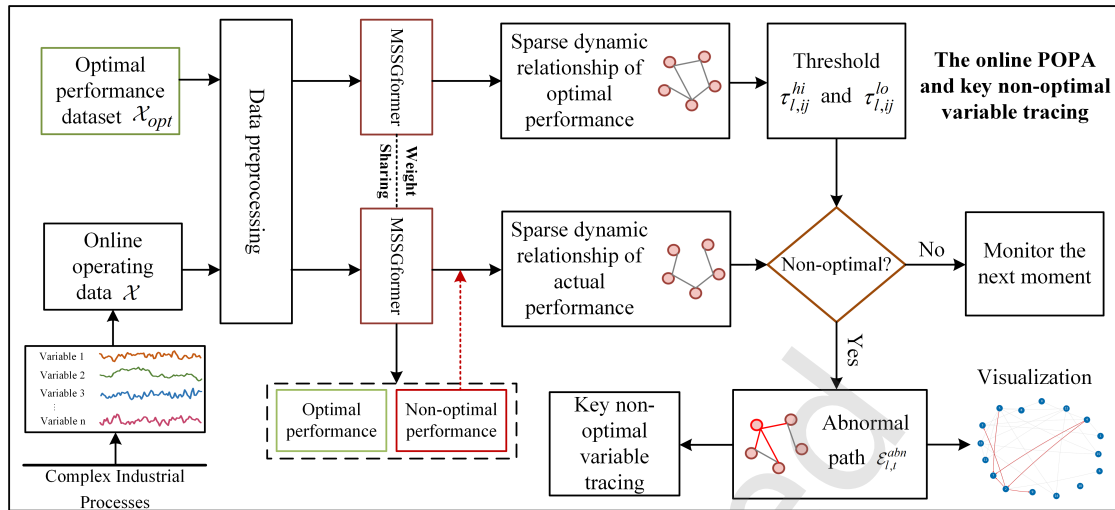


Fig. 3 Online POPA and key non-optimal variable tracing.

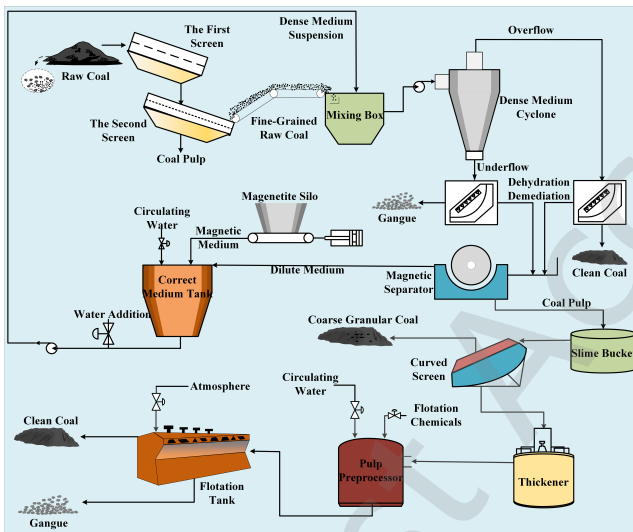


Fig. 4 Flow diagram of the coal preparation process [27].

objective of coal preparation is to effectively separate raw coal under complex operating conditions, remove impurities, and reduce ash content, thereby producing high-quality thermal coal that meets user requirements. Fig. 4 illustrates a typical coal preparation process [27], which mainly consists of three key subprocesses: the raw coal screening subprocess, the dense medium separation subprocess, and the coal slime flotation subprocess. Within each subprocess, multiple functional units are coupled through series-parallel structures. An abnormality in any critical unit may alter local separation performance and further propagate to downstream stages.

In complex industrial processes, operating performance is usually difficult to quantify directly and in real time using precise indicators. Therefore, abstracting operating performance into multi-level classification labels based on economic

performance helps improve operating performance discrimination and management convenience. In the coal preparation process, the final clean coal product mainly comes from the overflow product of the flotation cell. Therefore, its overflow ash content can effectively reflect the economic benefit of the overall process. A lower ash content indicates fewer impurities, higher product quality, and more significant corresponding economic returns. Therefore, this paper selects the flotation cell overflow ash content as the operating performance assessment indicator and divides the process operating performance into four levels, namely optimal, fine, medium, and poor, according to different degrees of deviation from the economically optimal operating performance level. These multi-level assessment labels are intuitive for operators, as shown in Table 1.

Meanwhile, according to process expert experience, 36 key monitoring variables are selected from the entire coal preparation process and divided into three subprocesses according to process functions. The specific variable distribution is shown in Tables 2 and 3. This partitioning strategy not only conforms to the process structure of coal preparation but also significantly reduces the complexity of modeling dynamic graph relationships. Specifically, if the 36 variables are globally modeled as a whole, the potential edge relationship scale is 1296. After partitioning them into three sub-blocks with 7, 13, and 16 variables, respectively, the intra-sub-block potential edge relationship scale becomes 474, which is 63.43% lower than that of global modeling. Therefore, multi-block modeling reduces the complexity of process modeling and improves the representation capability of local relationships. **Dataset partitioning and processing:** The experimental data were collected from a coal preparation process simulation

Table 1 Classification of operating performance levels and non-optimal variables in the coal preparation process

Clean coal ash content	Performance level	Non-optimal variable	Label
[5.01%, 6.00%)	Optimal	/	0
[6.00%, 7.20%)	Fine	Cyclone feed pressure	1
[7.20%, 8.41%)	Medium	Raw coal feed rate	2
[8.41%, 12.10%]	Poor	Flotation cell stirring speed	3

Table 2 Process variables of the coal preparation process

No.	Variable
1	Raw coal feed rate (kg/h)
2	Double-layer sieve underflow rate (kg/h)
3	Single-layer sieve overflow rate (kg/h)
4	Coal content in the mixture (%)
5	Mixing box discharge density (kg/m ³)
6	Mixing box discharge flow rate (m ³ /h)
7	Density of heavy medium entering the mixing box (kg/m ³)
8	Cyclone feed pressure (Pa)
9	Density of recovered heavy medium (kg/m ³)
10	Density of recovered heavy medium flow rate (m ³ /h)
11	Magnetic material addition amount (kg/h)
12	Volume flow rate of magnetic additive (m ³ /h)
13	Medium tank water flow rate (m ³ /h)
14	Medium tank discharge density (kg/m ³)
15	Medium tank liquid level (m)
16	Medium tank discharge flow rate (m ³ /h)
17	Cyclone overflow density (kg/m ³)
18	Cyclone underflow density (kg/m ³)
19	Cyclone overflow ash content (%)
20	Cyclone underflow ash content (%)
21	Coal slime bucket discharge mass rate (kg/h)
22	Coal slime bucket discharge flow rate (m ³ /h)
23	Coal slime bucket discharge density (kg/m ³)
24	Coal slime bucket liquid level (m)
25	Coal slime feed flow rate (m ³ /h)
26	Clean coal slime underflow flow rate (m ³ /h)
27	Clean coal slime underflow density (kg/m ³)
28	Thickener underflow density (kg/m ³)
29	Thickener underflow flow rate (m ³ /h)
30	Discharge density from slurry pretreatment unit (kg/m ³)
31	Discharge flow rate from slurry pretreatment unit (m ³ /h)
32	Flotation cell underflow ash content (%)
33	Flotation cell liquid level (m)
34	Flotation cell air intake rate (m ³ /m ² ·min)
35	Flotation cell stirring speed (rad/min)
36	Flotation reagent addition rate (g/L)

platform [27], with a sampling frequency of 1 Hz. Each raw sample contains a timestamp, 36 process variables, and the corresponding POPA indicator. A total of 19,000 continuous operating samples were collected in this study, and the dataset was strictly divided in chronological order without random shuffling. Specifically, the first 13,300 samples were used as offline training data, and the remaining 5,700 samples were

used as online testing data, with a split ratio of 7:3. The offline training data and online testing data were completely separated along the time axis. To ensure consistent comparison among all baseline models, the structures and basic training parameters of the baseline models were uniformly set before the experiments. No validation set was used for early stopping during training, and the testing set was used for final performance evaluation. The model input adopts a time window with a length of 100, and the training and testing windows were constructed independently within the training and testing sets, respectively, thereby avoiding data leakage caused by overlapping time windows or future information. The hyperparameter settings for model training are listed in Table 4.

The number of samples in each operating performance level is shown in Table 5. This paper does not adopt any class-balancing strategy because the level distribution conforms to the operating performance distribution characteristics of actual industrial processes, where optimal performance samples account for a relatively high proportion, followed by fine performance samples, while medium performance and poor performance samples are relatively fewer. To reduce the influence of class imbalance on the interpretation of assessment results, accuracy, weighted precision, weighted recall, and weighted F1-score are used for comprehensive model performance evaluation in the baseline model comparison.

To eliminate the influence of scale differences among different process variables on model training, z-score standardization is applied to the process variables:

$$x'_{j,t} = \frac{x_{j,t} - \mu_j^{train}}{\sigma_j^{train}} \quad (40)$$

where $x_{j,t}$ denotes the original value of the j -th process variable at the t -th sampling time, $x'_{j,t}$ denotes the standardized variable value, and μ_j^{train} and σ_j^{train} denote the mean and standard deviation of the j -th variable in the offline training set, respectively. The online testing set is standardized using the training-set statistics to avoid test information leakage.

4.2 Baseline Settings for Comparison

Baseline settings: MSSGformer is the model to be evaluated. MSSGformer-ATTN is a graph construction variant based on

Table 3 Sub-block partitioning of the coal preparation process

Sub-block ID	Variable indices	Node indices
Sub-block 1	1, 2, 3, 4, 5, 6, 7	0, 1, 2, 3, 4, 5, 6
Sub-block 2	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Sub-block 3	21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

Table 4 MSSGformer hyperparameter settings

Hyperparameter	Setting
Basic Parameters	
Batch size	10
Epoch	50
Learning rate	1×10^{-3}
Optimizer	Adam
Dropout rate	0.05
Input window length	100
Number of process variables	36
Number of classes	4
Model Parameters	
Number of blocks	3
Variables in block 1	7
Variables in block 2	13
Variables in block 3	16
GCN layers	2
GCN feed-forward dimension	64
λ_g	1.0
λ_{CVAE}	1.0
Number of attention heads	4
Encoder layers	2
Encoder feed-forward dimension	64

Table 5 Sample distribution of different operating performance levels

Dataset	Optimal	Fine	Medium	Poor
Training set	6496	3025	1834	1945
Testing set	2382	1430	923	965

the correlation adjacency matrix, while MSSGformer-EMM is a graph construction variant based on the expert-masked adjacency matrix. The baseline methods are summarized as follows:

AGCRN: Adaptive graph convolutional recurrent network, which combines adaptive graph learning and recurrent neural networks to model spatiotemporal dependencies [28].

ASTGAT: Adaptive spatiotemporal graph attention network, which combines the graph talking-heads attention layer and the gated temporal convolution layer to capture spatial dependencies and temporal dynamic features, respectively [29].

DSTAGNN: Dynamic spatiotemporal-aware graph neural network, which combines dynamic graph learning and spatiotemporal attention mechanisms to model dynamic spatiotemporal dependencies [30].

STGCN: Spatiotemporal graph convolutional network,

which combines graph convolution and one-dimensional convolution to model spatial and temporal dependencies [31].

GAE: Causal graph autoencoder network, which extends gradient-based continuous optimization for causal structure learning to the graph autoencoder framework [32].

CGSTAE: Causal graph spatiotemporal autoencoder network for process monitoring, which combines a correlation graph structure learning module based on the spatial self-attention mechanism with a spatiotemporal encoder-decoder module based on graph convolutional long short-term memory [17].

FEDformer: A model that combines seasonal-trend decomposition with Transformer to capture the global profile and detailed structure of time series, respectively [33].

Pyraformer: A model that uses a pyramidal attention structure to establish multi-scale temporal dependencies and capture temporal correlations over different ranges in a compact multi-resolution manner [34].

iTransformer: Inverted Transformer, which takes the variable dimension as the modeling subject to capture multivariate temporal dependencies [35].

Crossformer: A model that simultaneously models dependencies in the temporal and variable dimensions through segment-wise embedding and cross-dimension attention mechanisms [36].

Reformer: A model that uses locality-sensitive hashing attention to approximate standard self-attention computation, thereby improving the efficiency of long-sequence temporal modeling [37].

Informer: A model that selects key temporal dependencies through the ProbSparse attention mechanism to achieve efficient long-sequence dynamic modeling [38].

SPEDAE: Stacked performance-relevant enhanced denoising autoencoder network for POPA [39].

SSFALSTM: Supervised slow feature analysis LSTM network for POPA [11].

Training environment settings: The algorithms were implemented in Python 3.10.11 with PyTorch 2.1.1+cu121 and CUDA 12.6.65, and were run on a hardware platform equipped with an Intel® Core™ i5-13400F processor, an NVIDIA GeForce RTX 4060 Ti GPU, and 32 GB RAM.

Table 6 Benchmark comparison of POPA results

Model	Weighted Precision	Weighted Recall	Weighted F1-score	Accuracy
STGCN	82.05±3.35	84.40±2.33	82.68±2.91	84.40±2.33
ASTGAT	90.63±1.33	91.28±1.04	91.53±1.19	91.28±1.04
DSTAGNN	93.25±1.44	93.92±1.37	93.48±1.58	93.92±1.37
AGCRN	91.10±1.92	93.87±1.60	92.41±1.84	93.87±1.60
GAE	86.07±1.35	85.54±1.53	85.14±1.74	85.54±1.53
CGSTAE	92.00±1.21	89.76±1.55	90.44±1.24	89.76±1.55
FEDformer	85.58±0.97	84.91±0.94	83.94±1.05	84.91±0.94
Pyraformer	93.91±0.87	93.63±0.79	93.51±0.80	93.63±0.79
iTransformer	87.00±1.20	86.46±0.89	86.06±0.92	86.46±0.89
Crossformer	92.17±2.13	92.07±2.12	91.99±2.14	92.07±2.12
Reformer	94.67±0.80	94.47±0.78	94.37±0.81	94.47±0.78
Informer	95.68±0.33	95.60±0.35	95.60±0.35	95.60±0.35
SPEDAE	93.91±0.36	93.76±0.32	93.63±0.32	93.76±0.32
SSFALSTM	93.34±0.72	93.24±0.61	93.24±0.62	93.24±0.61
MSSGformer-ATTN	97.07±0.55	97.03±0.57	97.03±0.57	97.03±0.57
MSSGformer-EMM	96.91±0.78	96.88±0.78	96.87±0.78	96.88±0.78
MSSGformer	96.94±0.41	96.84±0.41	96.84±0.42	96.84±0.41

4.3 Topological Evolution Analysis of Graph Structure

This section quantitatively analyzes the topological evolution during graph structure construction to verify the pruning effect of the proposed method on weakly responsive redundant edges. Taking $l = 3$ and $t = 0$ as an example, Fig. 5 shows the evolution process from the initial correlation adjacency matrix to the final sparse dynamic adjacency matrix. This process contains three key stages: the correlation-based initial adjacency matrix A_l , the expert-masked adjacency matrix $\tilde{A}_l$, and the sparse dynamic adjacency matrix A_l^c obtained after sensitivity-based pruning with CVAE-PRM.

The evolution statistics of the number of edges within one batch are shown in Fig. 6. Taking the global block as an example, the correlation-based adjacency matrix A_{global} is 4.74×10^5 edges. Although this method can capture statistical correlations among variables, it often includes many redundant connections that violate physical transmission directions. Further, the expert masking matrix M_E imposes directional constraints on the correlation adjacency matrix and removes some edges that do not conform to process mechanisms, yielding the global masked adjacency matrix $\tilde{A}_{global}$ whose total number of edges is 3.26×10^5 . The results show that this step removes approximately 1.48×10^5 edges that do not conform to physical laws, with a reduction rate of 31.22%.

Based on this, CVAE-PRM further measures the sensitivity of dynamic edge relationships to local representation changes and selects perturbation-responsive edges, obtaining the final sparse dynamic adjacency matrix A_{global}^c . This stage produces the most decisive structural optimization, where the number of edges is sharply reduced from 3.26×10^5 to 3.12×10^4 , with a pruning rate of 90.43%. Combined with

the results of MSSGformer and its internal structural variants in Table 6, when the edge pruning rate reaches 93.42%, the assessment accuracy of MSSGformer decreases by only 0.19 percentage points compared with MSSGformer-ATTN, indicating that the overall assessment performance remains basically stable. This demonstrates that a large proportion of removed relationships have limited contributions to POPA, while the selected perturbation-responsive edges preserve the major operating performance discriminative information. Therefore, the edge relationship selection strategy based on CVAE-PRM can significantly reduce graph structural complexity while retaining valuable relationships for operating performance representation, thereby achieving dynamic graph sparsification for POPA.

4.4 POPA Experiments

POPA results: POPA aims to evaluate the degree to which the current performance deviates from the economically optimal operating performance based on process operating data. To comprehensively verify the assessment performance of MSSGformer, this paper selects multiple types of baseline methods for comparison, as described in Section 4.2.

Table 6 presents the POPA results of different methods on the source-domain testing set. All experiments were repeated five times, and the mean and standard deviation are reported. Precision, recall, and F1-score are all weighted metrics to reduce the influence of class imbalance on the interpretation of assessment results. The experimental results show that MSSGformer achieves 96.94%, 96.84%, 96.84%, and 96.84% in weighted precision, weighted recall, weighted F1-score, and accuracy, respectively, outperforming most

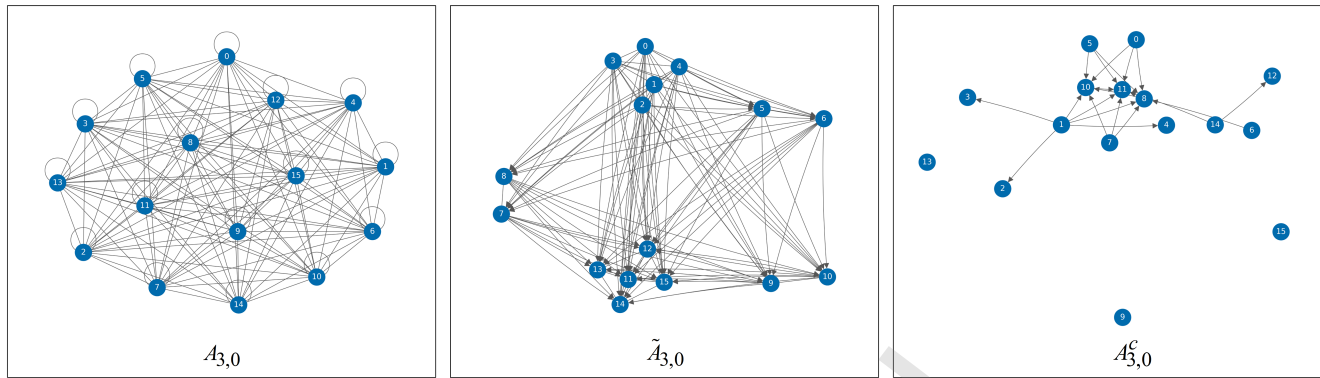


Fig. 5 Topological relationship evolution analysis.

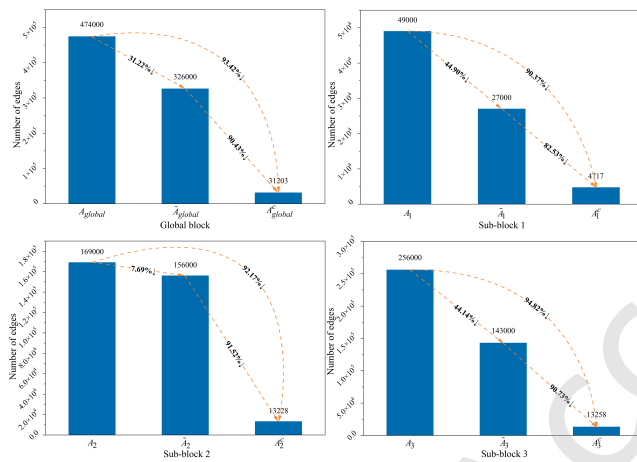


Fig. 6 Edge number evolution analysis.

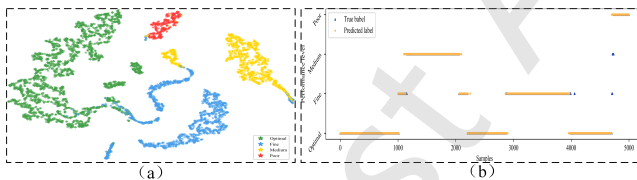


Fig. 7 Visualization of MSSGformer POPA results: (a) t-SNE feature visualization. (b) POPA results under different operating performance levels.

baseline models overall. Compared with Transformer-based methods that only model temporal dependencies, MSSGformer achieves higher assessment accuracy, indicating that relying solely on temporal modeling is insufficient to fully characterize multivariable coupling relationships in complex industrial processes. Meanwhile, compared with graph-based models that only focus on spatial relationships, MSSGformer jointly models the spatial dependencies among variables and temporal evolution features, achieving higher assessment accuracy. This demonstrates the key role of joint spatiotemporal modeling in improving POPA accuracy.

From the perspective of internal variants, MSSGformer-ATTN achieves an accuracy of 97.03% on the source domain,

slightly higher than the 96.84% of MSSGformer. This result indicates that although the complete correlation graph contains a large number of redundant edges, some of these redundant relationships may still contribute to the discrimination of source-domain samples. Therefore, the slight decrease of 0.19 percentage points in source-domain accuracy after substantial pruning in MSSGformer is acceptable. Combined with the graph structure evolution analysis in Section 4.3, MSSGformer still maintains assessment performance close to that of MSSGformer-ATTN after pruning 93.42% of graph edges. This indicates that the perturbation-responsive edges retained after CVAE-PRM pruning carry the major operating performance discriminative information, while the large number of pruned relationships make relatively limited contributions to overall POPA.

Fig. 7 further illustrates the feature representation effect of MSSGformer. The t-SNE visualization in Fig. 7(a) shows that different operating performance levels form relatively clear clustering structures in the embedding space, indicating that the model can learn process representations with operating performance discriminability. In Fig. 7(b), the predicted labels are generally consistent with the true labels, with deviations occurring only in a small number of boundary samples, further verifying the feasibility of MSSGformer for POPA in the coal preparation process.

Cross-distribution performance stability: To further evaluate the performance stability of MSSGformer under distribution shifts, this paper conducts validation from two aspects: cross-condition generalization capability and disturbance robustness. Two target-domain datasets are constructed accordingly. Target domain 1 is the operating condition shift dataset, where the main variations occur in the key variables listed in Table 1, including an increase in raw coal feed rate, abnormal changes in cyclone feed pressure, and changes in flotation cell stirring speed. This dataset is used to evaluate

the generalization capability of the model under unseen operating conditions. Target domain 2 is the noise perturbation dataset, where Gaussian noise with a standard deviation of 0.5 is added to the standardized process variables of the source-domain testing set. This dataset is used to evaluate the robustness of the model against measurement disturbances. Fig. 8 presents the Accuracy and performance degradation of different methods on the source domain and the two target domains.

The results show that MSSGformer achieves accuracies of 95.75% and 95.05% on target domain 1 and target domain 2, respectively, both of which are the best results among all methods. Meanwhile, when transferred from the source domain to target domain 1 and target domain 2, the accuracy drops of MSSGformer are 1.09 and 1.79 percentage points, respectively, which are significantly lower than those of most baseline methods. GAE and CGSTAE also show competitive generalization performance, mainly because their learned structures have a certain degree of stability. However, due to their relatively limited feature representation capability, their overall assessment performance remains limited. Therefore, considering both the target-domain evaluation accuracy and performance degradation, it can be observed that MSSGformer maintains high source-domain performance while achieving strong generalization capability under operating condition shifts and superior robustness against measurement noise.

Further comparison between MSSGformer and its internal variants shows that MSSGformer-ATTN achieves an accuracy of 97.03% on the source domain, but decreases to 92.96% and 92.51% on target domain 1 and target domain 2, respectively, with drops of 4.07 and 4.52 percentage points. MSSGformer-EMM shows improved generalization performance after introducing expert constraints. In contrast, MSSGformer maintains accuracies of 95.75% and 95.05% on target domain 1 and target domain 2, respectively, with drops of only 1.09 and 1.79 percentage points. Meanwhile, Fig. 9 shows the performance of different models in cross-condition POPA tasks. As the models gradually converge, MSSGformer maintains higher evaluation performance on target domain 1 while preserving high source-domain performance, whereas MSSGformer-ATTN and MSSGformer-EMM exhibit more significant performance gaps between the source domain and target domains.

The above results indicate that although correlation graphs can achieve slightly higher source-domain performance by utilizing more variable relationships, the involvement of a large number of candidate relationships in message propa-

gation also makes the model more susceptible to operating condition changes and input disturbances. MSSGformer employs CVAE-PRM to select correlation edges with significant responses to local representation changes, reducing the participation of low-value redundant relationships in information propagation. Therefore, the retained edges can carry critical operating performance discriminative information. Consequently, the proposed method reduces the interference caused by redundant relationships while improving cross-condition generalization capability and disturbance robustness.

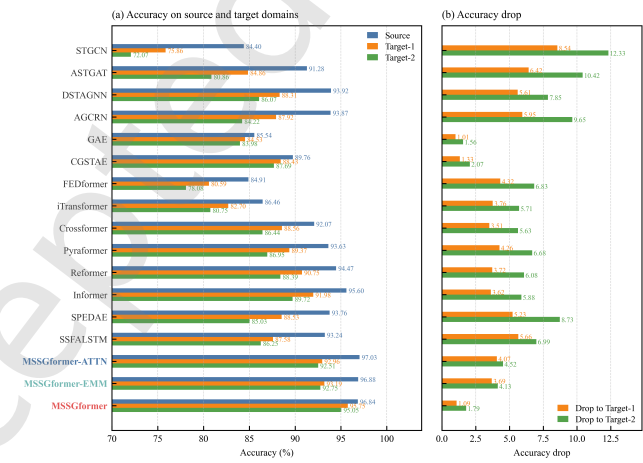


Fig. 8 Comparison of the generalization capability and disturbance robustness of different models. (a) Accuracy on the source domain and two target domains. (b) Accuracy degradation from the source domain to the two target domains.

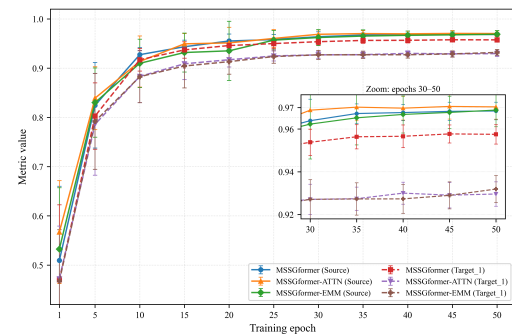


Fig. 9 Cross-condition POPA results of MSSGformer and its variants.

Table 7 Comparison of POPA accuracy between the removed-edge and retained-edge graphs

Graph type	Accuracy	Δ Acc over Removed-edge
Removed-edge	77.55 ± 0.89	12.37
Retained-edge	89.92 ± 0.76	

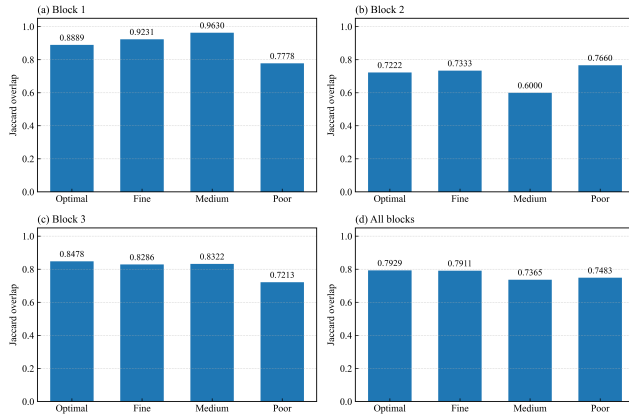


Fig. 10 Jaccard overlap of retained-edge sets under different operating performance levels.

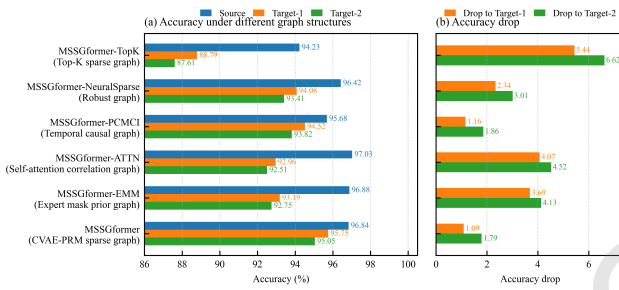


Fig. 11 SRRM ablation experiment.

4.5 Validation of Pruning Effectiveness and Structural Consistency

To further validate the effectiveness of the pruning results obtained by CVAE-PRM, this paper analyzes them from two perspectives: task contribution and structural consistency. It should be noted that, to avoid interference from the TRRM in operating performance representation, which may affect the evaluation of the contribution of the graph structure itself, MSSGformer only adopts the CVAE-PRM and POPAM in this experiment, so as to more directly verify the influence of graph structure changes on POPA performance. First, the retained-edge graph and the removed-edge graph are constructed and separately used for the POPA task to compare the contribution differences of the two types of edge relationships to operating performance representation. Specifically, the retained-edge graph consists of perturbation-responsive edges selected by CVAE-PRM, while the removed-edge graph consists of weakly responsive relationships eliminated by CVAE-PRM. Table 7 presents the assessment results of the retained-edge graph and the removed-edge graph on the POPA task. It can be seen that the accuracy of retained-edge is 89.92%, which is significantly higher than the 77.55% of removed-edge, with a difference of 12.37 percentage points. This result demonstrates that the perturbation-responsive

edges selected by CVAE-PRM contain more information beneficial for operating performance discrimination, whereas the pruned weakly responsive relationships provide relatively limited contributions to the POPA task. Therefore, the effectiveness of edge relationship selection is verified from the task perspective.

Furthermore, this paper evaluates the structural consistency of the pruning results obtained by CVAE-PRM under the same operating performance level by calculating the overlap degree of retained-edge sets. For each operating performance level, two groups of 20 samples are randomly selected, and the Jaccard similarity coefficient is adopted to measure the overlap between the two retained-edge sets:

$$J(E_a, E_b) = \frac{|E_a \cap E_b|}{|E_a \cup E_b|} \quad (41)$$

where E_a and E_b denote the two extracted testing groups, respectively.

As shown in Fig. 10, except for the overlap degree of 60.00% under the medium operating performance level in Block 2, the overlap degrees of core edges in most sub-blocks are higher than 70%. From the perspective of the overall graph structure, the overlap degrees of the optimal, fine, medium, and poor operating performance levels under All blocks are all higher than 73%, indicating that the retained edges have good structural consistency under the same operating performance. Combined with the higher POPA accuracy of retained-edge compared with removed-edge in Table 7, these results demonstrate that CVAE-PRM can not only select edge relationships with greater contributions to operating performance discrimination but also preserve relatively consistent structural patterns under the same operating performance level.

4.6 Module Analysis of MSSGformer

SRRM ablation experiment: The core function of SRRM is to construct a POPA-oriented sparse dynamic graph within each sub-block. To verify the effectiveness of the proposed graph pruning strategy, the graph construction method in MSSGformer is replaced with five alternative graph structures, including the Top-K sparse graph, NeuralSparse [40] robust graph, PCMCI [41] temporal causal graph, self-attention correlation graph, and expert-masked prior graph. The POPA performance of different graph structures on the source domain and target domains is compared, and the results are shown in Fig. 11.

The experimental results show that the MSSGformer-TopK sparse graph cannot guarantee that the retained relationships

possess high operating performance discriminative value, resulting in the poorest performance on both the source domain and target domains. The MSSGformer-NeuralSparse robust graph improves the performance on the source domain to some extent by learning robust structures and selecting effective connections, but still exhibits a significant performance degradation on the target domains. The MSSGformer-PCMC temporal causal graph achieves relatively stable performance across different operating conditions due to its strong structural stability, showing better generalization capability on the target domains. However, its relatively fixed structure and reliance on strict assumptions limit its task-specific adaptability, resulting in lower evaluation performance on both the source domain and target domains compared with the CVAE-PRM-based sparse graph structure.

Combined with the analysis in Section 4.4, MSSGformer-ATTN and MSSGformer-EMM utilize richer candidate relationships in the source domain and therefore achieve slightly higher source-domain evaluation accuracy, but exhibit more significant performance degradation under distribution shifts. MSSGformer measures the sensitivity of edge relationships to local representation changes through CVAE-PRM, selects perturbation-responsive edges that contribute to operating performance discrimination, and constructs a sparse dynamic graph. It achieves Accuracy values of 95.75% and 95.05% on target domain 1 and target domain 2, respectively, with performance degradations of only 1.09 and 1.79 percentage points. Overall, the sparse dynamic graph constructed based on the CVAE-PRM pruning strategy achieves a favorable balance between operating performance discriminative capability and cross-distribution performance stability.

TRRM ablation experiment: Considering that MSSGformer adopts a parallel spatiotemporal modeling structure, its TRRM is plug-and-play. Therefore, an ablation study is conducted on the TRRM module in MSSGformer to verify the ability of the framework to improve the performance of Transformer-based models. The Transformer variants embedded into MSSGformer are named as follows:

TRRM-FEDformer: MSSGformer embeds FEDformer as the TRRM.

TRRM-iTransformer: MSSGformer embeds iTransformer as the TRRM.

TRRM-Crossformer: MSSGformer embeds Crossformer as the TRRM.

TRRM-Pyraformer: MSSGformer embeds Pyraformer as the TRRM.

TRRM-Reformer: MSSGformer embeds Reformer as the TRRM.

TRRM-Informer: MSSGformer embeds Informer as the TRRM.

Fig. 12 compares the performance of different temporal modeling backbones when used independently and when embedded into MSSGformer. The results show that all temporal backbones achieve varying degrees of performance improvement after being embedded into MSSGformer. Specifically, the accuracies of TRRM-FEDformer, TRRM-iTransformer, and TRRM-Crossformer increase by 3.16, 2.85, and 3.40 percentage points, respectively. The accuracies of TRRM-Pyraformer, TRRM-Reformer, and TRRM-Informer increase by 1.67, 1.69, and 1.14 percentage points, respectively. These results indicate that temporal modeling alone is insufficient to fully capture the variable dependencies in complex industrial processes, while incorporating spatial relationship modeling through SRRM can significantly enhance the operating performance discriminative capability of POPA.

MDM ablation experiment: The masked decay matrix M is an important component in MSSGformer for characterizing temporal dependency features, and its hyperparameter configuration directly affects the model's perception strength of historical information, temporal propagation range, and the sensitivity of edge responses. To systematically evaluate the influence of different hyperparameter settings on model performance, this section conducts ablation experiments focusing on three key hyperparameters: the decay rate δ , the information propagation distance $d_{\max}$, and the perturbation threshold θ_l .

The decay rate δ is used to regulate the model's sensitivity to historical information, as shown in Fig. 13(a). When $\delta = 0$, historical information is completely masked, meaning that the model degenerates into a case without MDM constraints. As δ increases, the model becomes more capable of perceiving and retaining state features from historical time steps. Experimental results indicate that larger values of δ yield better POPA performance, whereas the model performance is the worst when $\delta = 0$. Accordingly, δ is set to 0.9.

The information propagation distance $d_{\max}$ determines the maximum effective temporal span of historical information that the model can perceive. Its physical interpretation corresponds to the effective memory length of historical process variables, as illustrated in Fig. 13(b). A small $d_{\max}$ leads to an overly narrow temporal window, making it difficult to capture rich historical contextual information, whereas an excessively large $d_{\max}$ may cause over-smoothing effects. Experimental results show that $d_{\max} = 40$ achieves favorable POPA performance, indicating that historical information within this range provides better interpretability for POPA.

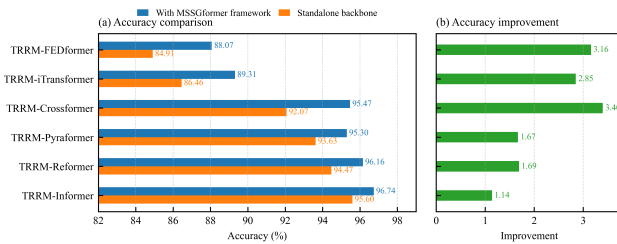


Fig. 12 TRRM ablation experiment.

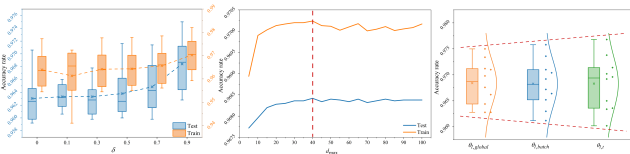


Fig. 13 Visualization of MDM ablation experiment.

The perturbation response threshold is used to select edges with significant responses. Since the operating conditions of complex industrial processes continuously evolve, the distribution of edge responses within different time windows may vary, making a fixed threshold difficult to adapt to different operating stages. Therefore, the perturbation response threshold is adaptively determined based on the statistical characteristics of the perturbation response matrix ΔA_t within the current input time window. During the training stage, CVAE-PRM calculates the corresponding perturbation response matrix and perturbation response threshold for each input window, and constructs sparse dynamic graphs for model optimization based on these results. During the online inference stage, the model parameters remain fixed. Only forward perturbation response calculation is performed for each newly input time window, and the corresponding perturbation response threshold and sparse dynamic graph are updated without backpropagation or parameter optimization. According to the current input time window, the perturbation response thresholds can be calculated at a single time step, batch time step, and global time step, denoted as $\theta_{l,t}$, $\theta_{l,batch}$, and $\theta_{l,global}$, respectively, as shown in Fig. 13(c). Experimental results demonstrate that $\theta_{l,global}$ provides the highest POPA accuracy while ensuring stable edge response selection and better adapting to process operating performance variations. Therefore, $\theta_l = \theta_{l,global}$ is adopted as the perturbation response threshold.

Key component ablation experiment: To analyze the contribution of key components in MSSGformer, we conduct ablation studies on the MSSGformer model to verify the contribution of each component to the performance improvement of the proposed model. The MSSGformer variants without different components are named as follows:

w/o CVAE: MSSGformer without the CVAE component, where the CVAE-PRM-based sensitivity pruning method is removed.

w/o SRRM: MSSGformer without the SRRM, where the spatial relationship representation module is removed.

w/o TRRM: MSSGformer without the TRRM, where the temporal relationship representation module is removed.

w/o EMM: MSSGformer without the EMM, where the expert masking matrix in the SRRM is removed.

w/o MDM: MSSGformer without the MDM, where the masked decay matrix in the SRRM and TRRM is removed.

Fig. 14 shows the POPA accuracy of different ablation variants. It can be seen that removing the TRRM reduces the model accuracy to 89.92%, resulting in the largest performance drop, which indicates that temporal relationship modeling is crucial for operating performance assessment in continuous industrial processes. After removing the SRRM, the accuracy decreases to 96.12%, indicating that spatial dependencies among variables also play an important role in operating performance representation. Removing the MDM reduces the accuracy to 96.30%, suggesting that temporal direction constraints and the decay mechanism help improve the utilization of historical information. After removing the EMM, the accuracy is 96.61%, indicating that process prior constraints can help reduce unreasonable connections.

It should be noted that w/o CVAE achieves an accuracy of 96.88% on the source domain, which is close to the 96.84% of the complete model. This indicates that the role of CVAE-PRM is not mainly reflected in improving source-domain accuracy, but rather in removing weakly responsive redundant relationships and enhancing the performance stability of the model under distribution shifts. Combined with the previous analysis, CVAE-PRM significantly reduces the number of graph edges while maintaining stable POPA performance and effectively decreases the performance degradation on target domains.

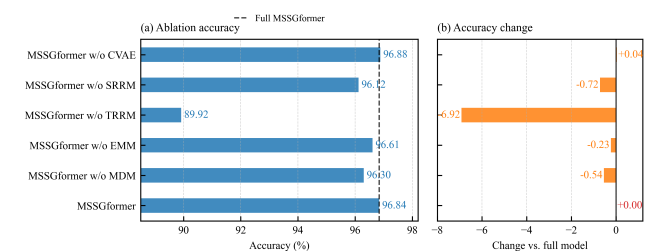


Fig. 14 Ablation experiments on different components of MSSGformer.

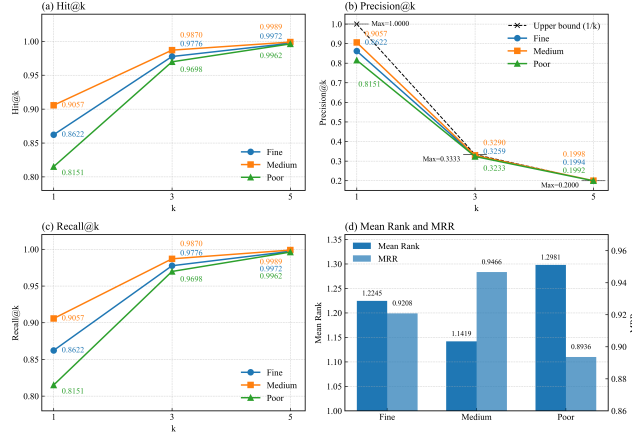


Fig. 15 Quantitative evaluation results of non-optimal variable localization performance.

4.7 Tracing of Key Non-optimal Variables

Changes in operating performance deviating from the economic optimum often induce variations in the strength of inter-variable structural relationships. Therefore, by analyzing the deviation of dynamic graph edge weights from the reference graph of optimal performance, the potential non-optimal variables can be traced, providing further evidence for operators to adjust process parameters and improve economic performance. When the edge relationship associated with a certain variable exhibits a significant deviation under the current operating performance level, it indicates that the variable may participate in the formation or propagation of non-optimal performance. Thus, it can be identified as a key non-optimal variable.

For quantitative evaluation, the tracing performance of MSSGformer is assessed using standard metrics from expert-annotated key non-optimal variable sets, including Hit@k, Precision@k, Recall@k, Mean Rank, and MRR. Since the Optimal operating performance level does not contain expert-labeled non-optimal variables, all metrics are computed only on the Fine, Medium, and Poor classes. Let the expert-labeled non-optimal variable set of the i -th test sample be G_i , and the Top- k variable set obtained according to the variable abnormality score be P_i^k , and the total number of test samples be N_{Test} .

The anomaly score of each variable is obtained based on the deviation of the sparse dynamic graph from the reference threshold, $Score(v)$ defined as:

$$s_{ij} = \max(A_{ij}^{c, test} - \tau_{ij}^{hi}, 0) + \max(\tau_{ij}^{lo} - A_{ij}^{c, test}, 0) \quad (42)$$

$$Score(v) = \sum_j s_{v,j} + \sum_i s_{i,v} \quad (43)$$

where $s_{v,j}$ denotes the abnormality score of the edge from variable v to variable j , $s_{i,v}$ denotes the abnormality score of the edge from variable i to variable v , with $j \neq v$ and $i \neq v$.

Hit@k measures whether any expert-labeled non-optimal variable appears in the Top-k predictions of the model, defined as:

$$\text{Hit@k} = \frac{1}{N_{Test}} \sum_{i=1}^{N_{Test}} \mathbb{I}(P_i^k \cap G_i \neq \emptyset) \quad (44)$$

where $\emptyset$ denotes the empty set and $\mathbb{I}(\cdot)$ is the indicator function, taking value 1 if the condition holds and 0 otherwise.

Precision@k measures the proportion of correctly identified non-optimal variables within the Top-k set:

$$\text{Precision@k} = \frac{1}{N_{Test}} \sum_{i=1}^{N_{Test}} \frac{|P_i^k \cap G_i|}{k} \quad (45)$$

Recall@k measures the proportion of expert-labeled non-optimal variables that are successfully retrieved:

$$\text{Recall@k} = \frac{1}{N_{Test}} \sum_{i=1}^{N_{Test}} \frac{|P_i^k \cap G_i|}{|G_i|} \quad (46)$$

Mean Rank measures the average ranking position of expert-labeled non-optimal variables in the predicted list:

$$\text{Mean Rank} = \frac{1}{N_{Test}} \sum_{i=1}^{N_{Test}} \text{rank}_i \quad (47)$$

MRR further emphasizes whether true non-optimal variables are ranked at top positions:

$$\text{MRR} = \frac{1}{N_{Test}} \sum_{i=1}^{N_{Test}} \frac{1}{\text{rank}_i} \quad (48)$$

Since each non-optimal operating performance level corresponds to only one expert-labeled non-optimal variable, the theoretical upper bound of Precision@k is $1/k$, and Recall@k is equivalent to Hit@k. Therefore, this work mainly adopts Hit@k, Mean Rank, and MRR as primary evaluation metrics, with Precision@k reported as a supplementary indicator.

The quantitative evaluation results of non-optimal variable localization performance are shown in Fig. 15. It can be seen that Hit@3 is higher than 0.96 and Hit@5 is higher than 0.99 for all three non-optimal operating performance levels, indicating that the model can stably rank the expert-labeled non-optimal variables within the top 3 to top 5 candidate variables. Meanwhile, the Mean Rank values of the three states are all close to 1, and the MRR values remain at a high level, indicating that the expert-labeled non-optimal variables are usually ranked at the top by the model. Although the observed values of Precision@3 and Precision@5 are relatively low, this is because each operating performance level contains only one true non-optimal variable, making

the theoretical upper bounds of Precision@k equal to 1/3 and 1/5, respectively. Therefore, the tracing results of the model are verified to be highly consistent with expert knowledge from the perspectives of Top-k hit rate and ranking quality.

Furthermore, Fig. 16 presents the tracing results of key non-optimal variables under four operating performance levels. For the optimal operating performance level, no obvious abnormal edge relationships are detected in each sub-block, and the overall dynamic graph structure remains stable, showing high consistency with expert experience and process knowledge. At the fine operating performance level, abnormal edges are mainly concentrated in sub-block 2. Among them, the edge weights between node 0 (cyclone feed pressure) and multiple nodes are significantly amplified, including node 1 (density of recovered heavy medium), node 4 (volume flow rate of magnetic additive), node 5 (medium tank water flow rate), and node 8 (medium tank discharge flow rate). As a preconditioning variable for sorting stability, fluctuations in the cyclone feed pressure disrupt the density balance of the media circulation system and trigger cascading changes in downstream operating quantities through loop coupling. These anomalous relationships indicate that the deviation of the fine operating performance level from the optimal operating performance level is mainly associated with fluctuations in cyclone feed pressure and their continuous transmission within the media circulation.

At the Medium operating performance level, abnormalities emerge earlier in sub-block 1, where node 0 (raw coal feed rate) exhibits significant abnormal edges. As the source variable of the entire coal preparation process, fluctuations in raw coal feed rate directly affect material distribution and concentration profiles in downstream sub-blocks. As shown in Fig. 16, this abnormality further propagates downstream to node 12 (cyclone underflow ash content) in sub-block 2. Underflow ash content is a critical indicator of separation quality, and its abnormality reflects degradation in separation precision caused by upstream feed disturbances. Meanwhile, in sub-block 3, node 2 (coal slime bucket discharge density) and node 7 (thickener underflow density) also exhibit pronounced anomalies, indicating that upstream material quality fluctuations ultimately lead to reduced stability of the flotation system.

At the Poor operating performance level, abnormalities are mainly concentrated in sub-block 3, particularly around node 14 (flotation cell stirring speed), which significantly deviates from the reference structure. Stirring speed is a key control variable governing bubble-particle attachment efficiency and froth stability. The sparse dynamic graph under

the Poor operating performance level shows that abnormalities are primarily triggered by internal faults within the flotation system, with weaker coupling to upstream sub-blocks, thereby exhibiting a localized abnormality pattern.

In summary, MSSGformer effectively distinguishes different operating performance levels and identifies potential key non-optimal variables by characterizing dynamic relationships in the sparse dynamic graph. These results are highly consistent with the process mechanisms and the variable relationships shown in Table 1, demonstrating that the proposed method can not only achieve accurate operating performance classification but also provide interpretable tracing results of key non-optimal variables, offering reliable support for precise diagnosis and operational decision-making in coal preparation processes.

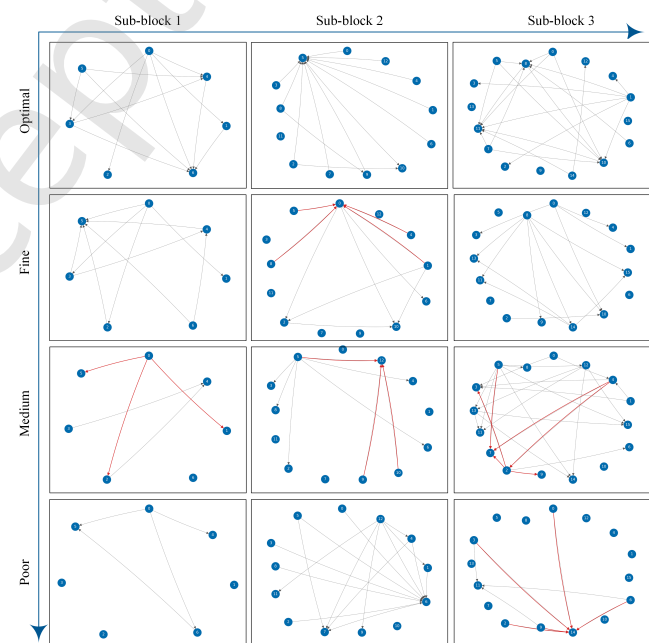


Fig. 16 Tracing results of key non-optimal variables under four operating performance levels.

5 Conclusion

To address the issue that numerous redundant spatiotemporal relationships interfere with operating performance representation and weaken the cross-distribution performance stability of models, this paper proposes an operating performance assessment method for complex industrial processes based on multi-block sparse spatiotemporal graph transformer. The proposed method achieves global operating performance assessment and key non-optimal variable tracing for complex industrial processes through sub-block-level spatiotemporal relationship modeling and process-level feature fusion.

At the spatial level, a CVAE-based perturbation response mechanism is designed to measure the sensitivity of dynamic edge relationships to local representation changes and select perturbation-responsive edges carrying critical operating performance discriminative information. Meanwhile, a masked decay matrix is proposed to transmit spatiotemporal information with temporal decay constraints, further enhancing the model's capability to represent spatiotemporal relationships in data. In addition, quantile thresholds are constructed based on the sparse dynamic graph selected under the optimal operating performance level, which are used to examine variations in edge relationship intensity and further trace key non-optimal variables. Experimental results demonstrate that MSSGformer maintains high operating performance assessment accuracy while effectively pruning a large number of weakly responsive relationships, achieving an effective trade-off between task relevance and cross-condition stability. The effectiveness and feasibility of the proposed method are verified in the operating performance assessment of the coal preparation process.

Future research will focus on predictive assessment for complex industrial processes, with comprehensive economic indicators serving as evaluation objectives to achieve forward-looking prediction of operating performance, thereby supporting early feedback and preventive maintenance and improving product quality and overall economic efficiency of industrial processes.

References

- [1] S. Li, B. Zhou, J. Shang, X. Chen, and J. Yu, "Review of recent applications and future perspectives on process monitoring approaches in industrial processes," *J. Manuf. Syst.*, vol. 82, pp. 509-530, Oct. 2025.
- [2] C. Zhang, W. Yin, K. Peng, and X. Zhang, "A deep semi-supervised echo state network-based distributed operating performance assessment framework for manufacturing processes optimized by enhanced black-winged kite algorithm," *Appl. Intell.*, vol. 55, no. 12, p. 834, Aug. 2025.
- [3] H. Zhang, Q. Liu, W. Fan, J. Shang, and L. Li, "Siamese Neural Network-based stationary feature extraction for nonstationary process monitoring," *Neurocomputing*, vol. 650, p. 130840, Oct. 2025.
- [4] Q. Jiang, X. Yan, and B. Huang, "Review and Perspectives of Data-Driven Distributed Monitoring for Industrial Plant-Wide Processes," *Ind. Eng. Chem. Res.*, vol. 58, no. 29, pp. 12899-12912, Jul. 2019.
- [5] K. Bu, Y. Liu, and F. Wang, "Operating performance assessment based on multi-source heterogeneous information with deep learning for smelting process of electro-fused magnesium furnace," *ISA Trans.*, vol. 128, pp. 357-371, Sep. 2022.
- [6] S. Wang, Y. Liu, L. Fu, F. Chu, F. Wang, and C. Bao, "Performance grade similarity-based generalized zero-shot operating performance assessment of industrial processes with insufficient samples," *J. Process Control*, vol. 154, p. 103523, Oct. 2025.
- [7] D. Lu, F. Wang, S. Wang, K. Bu, L. Zhang, and K. Li, "Operating Performance Assessment Based on Semi-Supervised Cluster Generative Adversarial Networks for Gold Flotation Process," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1-13, 2023.
- [8] R. Jia, M. Ren, J. Wang, F. Yu, and D. He, "Robust operating performance assessment of flotation processes using convolutional neural networks and feature learning," *Adv. Eng. Inf.*, vol. 64, p. 103087, Mar. 2025.
- [9] D. Hua, J. Dong, K. Peng, and S. Simani, "An Integrated Distributed Fault Diagnosis Framework for Large-Scale Industrial Processes Based on Spatio-Temporal Causal Analysis," *IEEE Trans. Ind. Inf.*, vol. 21, no. 8, pp. 6537-6546, Aug. 2025.
- [10] S. Xu, D. Pan, Z. Chen, Y. Fang, X. Qiu, and Z. Jiang, "Abnormal Condition Recognition of Blast Furnace Ironmaking Process Based on Time Series-Image Jointly Driven Deep Neural Network," *IEEE Trans. Ind. Inf.*, vol. 21, no. 8, pp. 6198-6209, Aug. 2025.
- [11] F. Chu, S. Liao, L. Hao, P. Wang, Y. Liu, and F. Wang, "Operating performance assessment method for industrial process with slowness principle-based LSTM network," *Eng. Appl. Artif. Intell.*, vol. 123, p. 106208, Aug. 2023.
- [12] H. Ren, X. Liang, C. Yang, Z. Chen, and W. Gui, "Spatio-temporal associations representation and application for process monitoring using graph convolution neural network," *Process Saf. Environ. Prot.*, vol. 180, pp. 35-47, Dec. 2023.
- [13] M. Jia, Y. Yao, and Y. Liu, "Review on Graph Neural Networks for Process Soft Sensor Development, Fault Diagnosis, and Process Monitoring," *Ind. Eng. Chem. Res.*, vol. 64, no. 17, pp. 8543-8564, Apr. 2025.
- [14] D. Lu, F. Wang, Y. Liu, S. Wang, and K. Li, "Process Operating Performance Assessment for Plant-Wide Froth Flotation via Distributed Multigraph Deep Embedding Graph Clustering Network," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1-10, 2024.
- [15] L. Hao, F. Chu, T. Chen, M. Jia, and F. Wang, "Operating performance assessment of industrial process based on MIC-graph convolutional networks with local slow feature analysis," *Neural Networks*, vol. 191, p. 107773, Nov. 2025.
- [16] C. Duan, Z. Wu, L. Zhu, X. Xu, J. Zhu, Z. Wei, and X. Yang, "DSTN: Dynamic Spatio-Temporal Network for Early Fault Warning in Chemical Processes," *Knowledge-Based Syst.*, vol. 296, p. 111892, Jul. 2024.
- [17] X. Zhang, C. Song, W. Dai, Z. Zhang, K. Gao, and F. Gao, "Causal Graph Spatial-Temporal Autoencoder for Reliable and Interpretable Process Monitoring," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1-13, 2026.
- [18] C. Dong, D. Li, and H. R. Karimi, "Reinforcement learning driven adaptive graph construction for fault diagnosis of chem-



- ical processes,” *Reliab. Eng. Syst. Saf.*, vol. 266, p. 111781, Feb. 2022.
- [19] D. Li, Y. Bao, and X. Xu, “Enhancing Graph Reconstruction: Uniting Dual-Level Graph Structure With Graph Reinforcement Learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 10, pp. 18218–18228, Oct. 2025.
- [20] H. Liu, J. Wu, W. Li, and X. Yu, “Graph structure learning with joint node and structural feature representation for node classification,” *Neurocomputing*, vol. 663, p. 132002, Jan. 2026.
- [21] Y. Wang, Y. Xu, J. Yang, M. Wu, X. Li, L. Xie, and Z. Chen, “Fully-Connected Spatial-Temporal Graph for Multivariate Time-Series Data,” in: *Proc. AAAI Conf. Artif. Intell.*, AAAI, 2024.
- [22] A. Shehzad, F. Xia, S. Abid, C. Peng, S. Yu, D. Zhang, and K. Verspoor, “Graph Transformers: A Survey,” *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–20, Jan. 2026.
- [23] T. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” in: *Int. Conf. Learn. Represent.*, ICLR, 2017.
- [24] W. Wu, C. Song, J. Zhao, Z. Xu, and W. Yu, “Hybrid-Order Graph Embedded Distributed Encoder-Decoder for Multiunit Industrial Plant-Wide Process Monitoring,” *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 4, pp. 7298–7311, Oct. 2024.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in: *Adv. Neural Inform. Process. Syst.*, NeurIPS, 2017.
- [26] C. Gong, C. Zhang, D. Yao, J. Bi, W. Li, and Y. Xu, “Causal Discovery from Temporal Data: An Overview and New Perspectives,” *ACM Comput. Surv.*, vol. 57, no. 4, Dec. 2024.
- [27] F. Chu, H. Feng, D. Zou, K. Li, and T. Song, “Building Block Modeling Method for Complex Industrial Processes Based on Federated Learning,” *IEEE Internet Things J.*, vol. 12, no. 22, pp. 46971–46984, Nov. 2025.
- [28] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, “Adaptive graph convolutional recurrent network for traffic forecasting,” in: *Adv. Neural Inform. Process. Syst.*, NeurIPS, 2020.
- [29] X. Kong, J. Zhang, X. Wei, W. Xing, and W. Lu, “Adaptive spatial-temporal graph attention networks for traffic flow forecasting,” *Appl. Intell.*, vol. 52, no. 4, pp. 4300–4316, Mar. 2022.
- [30] S. Lan, Y. Ma, W. Huang, W. Wang, H. Yang, and P. Li, “DSTAGNN: Dynamic Spatial-Temporal Aware Graph Neural Network for Traffic Flow Forecasting,” in: *Int. Conf. Mach. Learn.*, ICML, 2022.
- [31] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” in: *Proc. AAAI Conf. Artif. Intell.*, AAAI, 2018.
- [32] I. Ng, S. Zhu, Z. Chen, and Z. Fang, “A Graph Autoencoder Approach to Causal Structure Learning,” in: *Adv. Neural Inform. Process. Syst.*, NeurIPS, 2019.
- [33] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, “FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting,” in: *Int. Conf. Mach. Learn.*, ICML, 2022.
- [34] S. Liu, H. Yu, C. Liao, J. Li, W. Lin, A. Liu, S. Dustdar, “Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting,” in: *Int. Conf. Learn. Represent.*, ICLR, 2022.
- [35] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, “iTransformer: Inverted Transformers Are Effective for Time Series Forecasting,” in: *Int. Conf. Learn. Represent.*, ICLR, 2024.
- [36] Y. Zhang and J. Yan, “Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting,” in: *Int. Conf. Learn. Represent.*, ICLR, 2023.
- [37] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The Efficient Transformer,” in: *Int. Conf. Learn. Represent.*, ICLR, 2020.
- [38] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting,” in: *Proc. AAAI Conf. Artif. Intell.*, AAAI, 2021.
- [39] Y. Liu, Z. Ma, F. Wang, R. Ma, F. Chu, X. Li, and C. Guan, “Operating performance assessment based on stacked performance-relevant enhanced denoising auto-encoder for industrial processes,” *Can. J. Chem. Eng.*, vol. 102, no. 4, pp. 1509–1521, Apr. 2024.
- [40] C. Zheng, B. Zong, W. Cheng, D. Song, J. Ni, W. Yu, H. Chen, and W. Wang, “Robust graph representation learning via neural sparsification,” in: *Int. Conf. Mach. Learn.*, ICML, 2020.
- [41] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, “Detecting and quantifying causal associations in large nonlinear time series datasets,” *Sci. Adv.*, vol. 5, no. 11, Nov. 2019.

Author biography



Ruicong Zhang received the B.S. degree in information and computing Science from Huaiyin Normal University, Huai'an, China, in 2020, and the M.S. degree in information security from China University of Mining and Technology, Xuzhou, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China.

His current research interests include fault diagnosis, process operating performance assessment, energy optimization of wireless sensors, and time-frequency analysis, computer vision.



Fei Chu (Senior Member, IEEE) received the B.S. degree in electrical and automation from Qingdao University, Qingdao, China, in 2007, and the M.S. and Ph.D. degrees in control theory and control engineering from Northeastern University, Shenyang, China, in 2009 and 2014, respectively. He is currently a

Professor with the School of Information and Control Engineering, the Ministry of Education Engineering Research Center of Intelligent Control for Underground Space, the Artificial intelligence research institute, and the Xuzhou Key Laboratory of Artificial Intelligence and Big Data, China University of Mining and Technology, Xuzhou, China.

His current research interests include modeling, control, and optimization of complex industrial process, machine learning, statistical process monitoring, and process operating performance assessment.



Shuyan Xu received the B.S. degree in Electrical Engineering and Automation from Soochow University, Suzhou, China, in 2023. He is currently pursuing the M.S. degree with the College of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China.

His current research interests include fault diagnosis, process operating performance assessment, and critical transition of operating conditions.



Fuli Wang received the B.E., M.E., and Ph.D. degrees from Northeastern University, Shenyang, China, in 1982, 1985, and 1988, respectively. He is currently a Professor with the College of Information Science and Engineering, Northeastern University.

His main research interests are modeling, control, and optimization for complex systems, monitoring and health maintenance for industrial systems, artificial intelligence, and fault diagnosis.