Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3627-3636
Journal homepage: http://iieta.org/journals/ts

Terminal Abnormal Behavior Detection and Spatiotemporal Interpretability Analysis Based on Multimodal Visual-Behavioral Joint Representation

Li Di[1*], Lijing Yan[2], Yifan Song[2], Han Liu[2], Minghao Dong[3]

[1] State Grid Henan Electric Power Company, Zhengzhou 450000, China
[2] State Grid Henan Information & Telecommunication Company, Zhengzhou 450000, China
[3] State Grid Henan Electric Power Company Zhumadian Power Supply Company, Zhumadian 450000, China

Corresponding Author Email: edwarddi@outlook.com

https://doi.org/10.18280/ts.420644

Received: 30 April 2025
Revised: 19 September 2025
Accepted: 12 October 2025
Available online: 31 December 2025

Keywords:
IIoT terminals, abnormal behavior detection, visual guidance, Spatiotemporal Graph Attention Network (ST-GAT), multimodal fusion, spatiotemporal interpretability, behavioral understanding

ABSTRACT

The security of industrial Internet of Things (IIoT) terminals relies on the integration of visual monitoring and operational telemetry technologies, but this process faces core challenges, such as inefficient heterogeneous modality fusion and inadequate abnormal reasoning interpretability. Abnormal behaviors in industrial terminals are characterized by both behavioral sequence ambiguity and visual state correlation. Single-modality detection can easily lead to misjudgments, making a visual-guided multimodal fusion breakthrough urgently needed. Existing methods have three main limitations: visual and behavioral modalities often employ static concatenation without utilizing visual information to decouple behavioral ambiguity; dynamic spatiotemporal dependency modeling lacks theoretical support from visual guidance, and weight allocation is highly subjective; and interpretability is limited to a single dimension, lacking a closed-loop system of visual evidence, behavioral logic, and causal traceability. In response, this paper proposes a Visual-guided Multimodal Spatiotemporal Graph Attention Network (VG-MS-ST-GAT), which achieves high-accuracy detection and deep interpretability through four core modules: the visual-behavioral spatiotemporal dynamic interaction module guides through visual features of device regions of interest (ROI) and operator actions, generating spatiotemporal graph dynamic weights using a small multilayer perceptron to model cross-modal spatiotemporal dependencies; the causal-guided intent feature extraction module formalizes intent as latent variables linking multimodal sequences to abnormal states, and uses cross-modal attention and causal decoupling mechanisms to extract fine-grained intent representations; the cross-modal intent recognition module constructs a visual-behavioral contrastive learning loss to enhance the distinguishability of intent features for abnormal classification; and the virtual-physical interactive interpretability output module integrates attention heatmaps with Granger causality tests to provide multidimensional explanations, including visual anomaly regions, key behavioral sequences, and causal propagation paths. The core contributions of this study include: proposing a visual-behavioral alignment-based spatiotemporal graph modeling paradigm, using visual-guided dynamic weight generation to address the challenges of asynchronous and heterogeneous multimodal fusion; establishing a causal intervention-driven intent decoupling representation mechanism, capturing subtle abnormal precursors through visual precursors and behavioral logic; constructing a virtual-physical interactive interpretability framework for operations and maintenance, transforming model decisions into actionable traceability reports that include visual evidence, behavioral links, and causal roots; and building the Ind-ViBe-2024 dataset, which contains 10 types of terminals, 8 types of anomalies, and 52,000 multimodal samples, providing a benchmark testing platform for industrial visual-behavioral abnormal detection.

1. INTRODUCTION

The deepening of Industry 4.0 has made industrial terminals the core hub of smart manufacturing systems [1, 2], with their stable operation directly determining production efficiency and system security [3]. However, terminal abnormal behaviors, such as equipment failures, operator violations, and communication protocol attacks [4, 5], have caused more than $300 billion in global industrial economic losses annually.

In industrial scenarios, over 70% of terminal abnormalities manifest as both visual state changes and behavioral sequence shifts. However, existing detection solutions face significant modal fragmentation issues: visual-based methods can only capture physical state anomalies and fail to relate to the logical correlation of operational instructions [6, 7]; behavioral-based methods struggle to distinguish normal operations from malicious behaviors in ambiguous scenarios [8, 9]. The core value of image processing technologies lies in encoding fine-grained visual features such as the color of equipment indicator lights, component deformations, and temporal visual features such as operator button gestures and tool usage postures. These features effectively decouple the ambiguity of behavioral modalities, which is key to improving the robustness of anomaly detection and forms an interdisciplinary innovation direction that top image processing journals focus on.

Industrial terminal anomaly detection technology has evolved from early rule-based methods to single-modality data-driven solutions and is currently moving towards multimodal fusion-driven approaches [10]. Among them, the paradigm for processing visual modalities has evolved from static image feature extraction to sequential video feature encoding, ultimately forming a spatiotemporal visual representation technical route, with spatiotemporal graph neural networks becoming the mainstream tool for modeling the spatial correlation and temporal dependencies of multiple terminals [11, 12]. However, existing research still has multiple-dimensional limitations: firstly, the fusion of visual and behavioral modalities mostly adopts static feature concatenation, without dynamically adjusting the weight allocation of behavioral features based on visual information, leading to the inability to focus on behavior sequences when equipment visual state anomalies occur [13, 14]; secondly, visual spatiotemporal modeling has not adapted to the uniqueness of industrial scenarios and lacks specialized feature extraction mechanisms for issues such as device key region ROI, lighting changes, and occlusion [15, 16]; thirdly, interpretability solutions can only output attention heatmaps of behavior sequences, without forming a “visual evidence-behavior logic” interactive explanation by combining visual anomaly regions, making it difficult to gain the trust of operations and maintenance personnel [17, 18]; fourthly, existing datasets mostly consist of single visual or behavioral data, lacking aligned labels for “device visual state-operational behavior-abnormal root cause” in industrial scenarios, limiting the training and validation of multimodal methods [19, 20].

The research goal of this paper is to propose a visual-guided multimodal spatiotemporal fusion method to solve the core problems in industrial terminal anomaly detection, such as "imprecise modality fusion, inadequate visual modeling, and lack of evidence chains in explanations," while meeting the dual requirements of top journals for image processing innovation and theoretical depth. Specific core contributions include: firstly, proposing a visual-guided dynamic spatiotemporal graph construction method—relying on device visual features to adaptively generate the adjacency matrix weights of the spatiotemporal graph, overcoming the limitations of traditional static weights that cannot match dynamic changes in working conditions, and providing visual-driven theoretical support for the efficient fusion of multimodal features; secondly, designing a cross-modal visual-behavioral attention mechanism—clearly focusing the visual branch on device state changes and spatial correlations, and the behavioral branch on operational logic and temporal dependencies, achieving precise alignment of key information between modalities through cross-query; thirdly, establishing a causal-enhanced virtual-physical interactive interpretability framework—integrating visual attention heatmaps with Granger causality tests to generate a traceability path of "visual anomaly areas-associated terminals-behavioral abnormal sequences," achieving the adaptation of explanation results to the industrial operations and maintenance knowledge system; fourthly, constructing the industrial visual-behavioral aligned dataset Ind-ViBe-2024—which contains 52,000 multimodal samples from 10 types of industrial terminals, annotating visual ROI features, operational behavior sequences, and abnormal causal chains, filling the gap in industrial multimodal aligned data.

The structure of this paper is as follows: Chapter 2 reviews the research progress in related fields, identifying the core gaps in existing methods; Chapter 3 elaborates on the visual-guided multimodal model architecture and key module designs; Chapter 4 validates the effectiveness and robustness of the method through experiments enhancing visual features; Chapter 5 discusses the engineering value and limitations of the experimental results; Chapter 6 summarizes the paper and looks forward to future research directions.

2. METHOD

2.1 Problem definition

Consider a system composed of N industrial terminals, where the set of terminals is denoted as $T = \{T_1, T_2, \dots, T_N\}$. The multimodal input for each terminal is defined as $X_i = \{V_i, B_i\}$, where $V_i \in R^{H\times W\times C\times T}$ represents the spatiotemporal visual information of the terminal: H and W are the height and width of the device status image, C is the number of image channels, and T is the time step. The content includes both static features of the device's key regions and the temporal video frames of the operator's actions. The behavioral modality $B_i \in R^{D\times T}$ represents the temporal behavioral features of the terminal, where D is the dimension of behavioral features, including operational instructions, communication frequency, operating parameters, and other data. The visual and behavioral modalities must be strictly synchronized along the time dimension to provide the basis for cross-modal spatiotemporal correlation modeling.

To accurately characterize the intrinsic relationship between multimodal observation sequences and terminal abnormal states, "intent" is formalized as a latent variable connecting the two, denoted as $I_i(t) \in R^K$, satisfying the mapping relationship $I_i(t)=f(V_i(t),B_i(t))$, where K is the intent feature dimension and $f(\ )$ is the multimodal feature fusion function. The normal intent space is defined as Ω; when $I_i(t) \in \Omega$, the terminal is in normal operating status, and when $I_i(t) \notin \Omega$, the terminal exhibits abnormal behavior. Based on this, the goal of the task is as follows: given the multimodal inputs $X = \{X_1, \dots, X_N\}$for the first T time steps, predict the terminal state $Y = \{y_1, \dots, y_N\}$ at the $T + 1$ time step, where $y_i = 0$ indicates normal and $y_i = 1$ indicates abnormal; and simultaneously output the interpretability results $E = \{E_v, E_t, E_c\}$, where E_v is the visual abnormality evidence, E_t is the abnormal time-series link, and E_c is the cross-terminal abnormal causal relationship.

To comprehensively evaluate the performance of the method, two types of metrics are used to quantify detection accuracy and interpretability. Classification metrics include accuracy (ACC), precision, recall, F1-score, and AUC-ROC, which measure the overall effectiveness of abnormal detection from different dimensions. Interpretability metrics consist of three core indicators: the visual-behavioral attention consistency C, which measures the matching degree of attention weights between visual anomaly areas and behavioral abnormal sequences; the anomaly traceability accuracy A, which evaluates the alignment between the model's output traceability path and the real abnormal causal chain; and the explanation fidelity F, which quantifies the correlation between the explanation results and the model's decision logic. The calculation formula for F is:

$$F=1-\frac{|p-p'|}{max\,(p,1-p)} \tag{1}$$

where, p is the original abnormal prediction probability, p' is the prediction probability after perturbing the visual or behavioral key areas, and a higher value of F indicates that the explanation result is more faithful to the model's actual reasoning process.

2.2 Overall model framework

The VG-MS-ST-GAT proposed in this paper adopts a hierarchical collaborative architecture. The overall process is as follows: input layer → visual-behavioral spatiotemporal dynamic interaction module → causal-guided intent feature extraction module → cross-modal intent recognition module → virtual-physical interactive interpretability output module → output layer. Figure 1 shows the overall architecture of the VG-MS-ST-GAT model proposed in this paper, with the raw visual and behavioral features of industrial terminals as inputs. The input layer performs targeted preprocessing of multimodal data: the visual modality extracts key region features and encodes spatiotemporal information to obtain visual features rich in device status and operational actions, while the behavioral modality is standardized and encoded to generate structured operational logic features. The preprocessed bimodal features enter the visual-behavioral spatiotemporal dynamic interaction module, where visual features guide the construction of dynamic spatiotemporal graphs and the allocation of cross-modal attention, achieving precise alignment of heterogeneous features. The subsequent causal-guided intent feature extraction module separates modal noise through a causal decoupling mechanism and extracts fine-grained intent representations from the associated features. The cross-modal intent recognition module performs anomaly classification based on this representation, while driving the virtual-physical interactive interpretability output module to generate visual anomaly evidence, abnormal temporal sequences, and cross-terminal causal relationships. Finally, the output layer integrates anomaly detection results and a multidimensional interpretability report, forming a closed-loop reasoning system of "feature fusion - intent recognition - decision explanation."

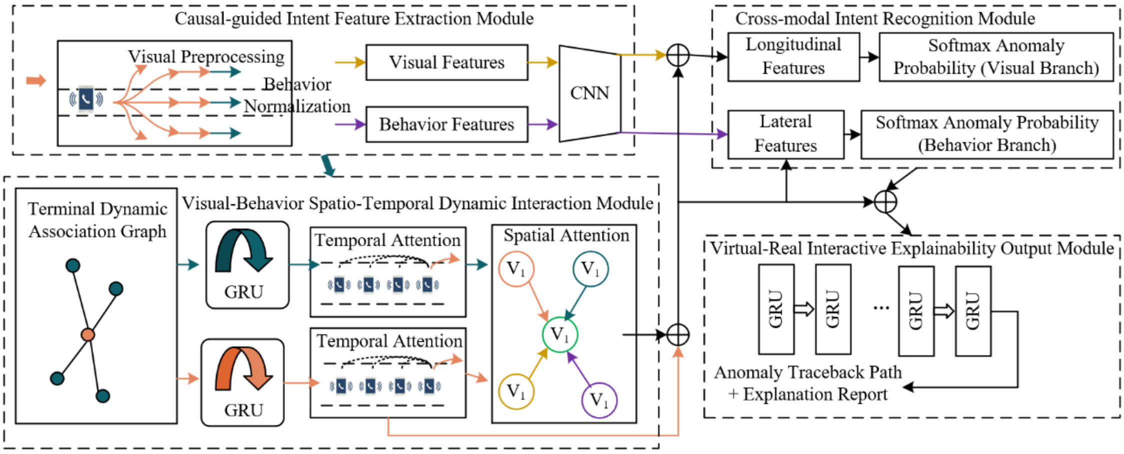

Causal-guided Intent Feature Extraction Module
Visual Preprocessing
Behavior Normalization
Visual Features
Behavior Features
CNN
Cross-modal Intent Recognition Module
Longitudinal Features
Softmax Anomaly Probability (Visual Branch)
Lateral Features
Softmax Anomaly Probability (Behavior Branch)
Terminal Dynamic Association Graph
Visual-Behavior Spatio-Temporal Dynamic Interaction Module
GRU
GRU
Temporal Attention
Temporal Attention
Spatial Attention
V1
V1
V1
V1
V1
Virtual-Real Interactive Explainability Output Module
GRU
GRU
...
GRU
GRU
Anomaly Traceback Path
+ Explanation Report

Figure 1. VG-MS-ST-GAT model structure diagram

2.3 Visual-behavioral spatiotemporal dynamic interaction module

The core objective of the visual-behavioral spatiotemporal dynamic interaction module is to generate fusion features containing spatiotemporal information and modality synergy characteristics through industrial visual adaptation processing and cross-modal correlation modeling, providing high-quality input for subsequent intent extraction. Given the characteristics of industrial visual data such as lighting fluctuations, frequent occlusions, and dispersed key information, the module first performs a three-step preprocessing process: it uses YOLOv8 to locate and crop key regions such as device control panels and indicator lights, focusing on core visual information; it applies adaptive histogram equalization to eliminate lighting differences, and combines generative adversarial networks to complete occluded regions, thereby enhancing data robustness; finally, it extracts spatial texture features using the lightweight convolutional neural network EfficientNet-B0, paired with ConvLSTM to capture the temporal dependencies of operator actions and device statuses, outputting visual features with integrated spatiotemporal information, denoted as $h_i^v \in R^{K_v \times T}$, where K_v is the visual feature dimension, and T is the time step. The behavioral modality is then standardized and encoded to generate structured behavioral features h_i^b, which are synchronized with the visual features in time.

Based on the preprocessed bimodal features, the module constructs a dynamic temporal graph $G_t = (V, E_t)$to model the dynamic associations among multiple terminals. The node set V is composed of the concatenated visual and behavioral features of each terminal, and the adjacency matrix $E_t \in R^{N\times N}$ characterizes the real-time association strength between terminals. To overcome the limitations of traditional manually set weights, the module uses a multilayer perceptron to dynamically generate weight coefficients $[\alpha_{i,j}(t),\beta_{i,j}(t),\gamma_{i,j}(t)]=MLP(h_i^v(t),h_j^v(t))$, which are combined with the physical distance $Dist(T_i,T_j)$, communication frequency $Comm(T_i,T_j)$, and operational synergy $Op(T_i,T_j)$ to construct the adjacency matrix:

$$\varepsilon_t(i,j)=\alpha_{i,j}(t)\cdot Dist(T_i,T_j)+\beta_{i,j}(t)\cdot Comm(T_i,T_j)+\gamma_{i,j}(t)\cdot Op(T_i,T_j) \tag{2}$$

Figure 2 shows the local schematic diagram of the visual-guided dynamic spatio-temporal graph attention model constructed, where the nodes correspond to typical terminals in the industrial scenario, and the edge weights are generated by calculating the similarity of visual features between terminals. It intuitively demonstrates the dynamic graph construction logic of "visual feature-guided terminal association," where the weight of self-association edges is reinforced by behavior features. This design allows the terminal association strength to dynamically adjust according to the visual state. For example, when a visual anomaly occurs in the device, the communication association weight with related terminals can be automatically increased, achieving precise adaptation of association modeling to operational conditions.

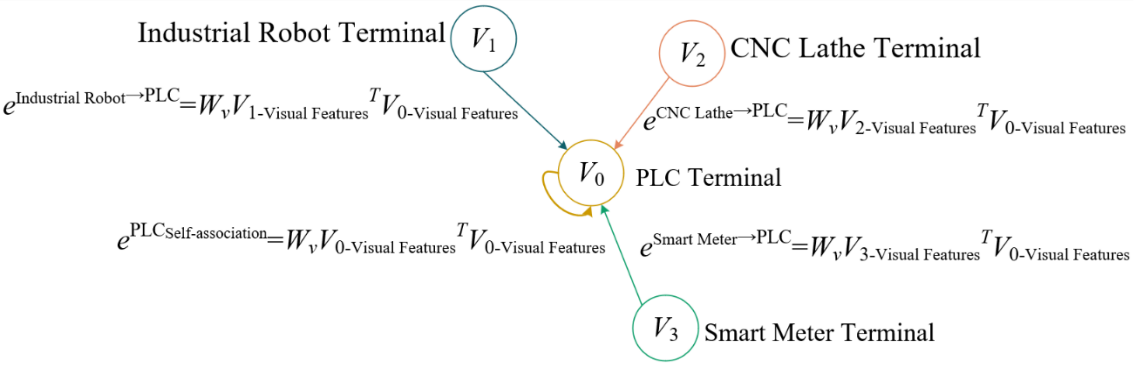

Industrial Robot Terminal
V_1
V_2
CNC Lathe Terminal
$e^{\text{Industrial Robot}\rightarrow\text{PLC}}=W_v V_{1\text{-Visual Features}}{}^T V_{0\text{-Visual Features}}$
$e^{\text{CNC Lathe}\rightarrow\text{PLC}}=W_v V_{2\text{-Visual Features}}{}^T V_{0\text{-Visual Features}}$
V_0
PLC Terminal
$e^{\text{PLCSelf-association}}=W_v V_{0\text{-Visual Features}}{}^T V_{0\text{-Visual Features}}$
$e^{\text{Smart Meter}\rightarrow\text{PLC}}=W_v V_{3\text{-Visual Features}}{}^T V_{0\text{-Visual Features}}$
V_3
Smart Meter Terminal

Figure 2. Schematic of the VG-MS-ST-GAT model

To achieve precise alignment of key information in both modalities, the module further designs a dual-branch attention structure and cross-modal collaborative mechanism. The visual branch focuses on the device's visual state changes and spatial associations. By concatenating the visual features of terminal i and j, $[h_i^v\|h_j^v]$, and mapping them through the weight matrix W_v and activation, the attention score a_{ij}^v is generated as $a_{ij}^v=LeakyReLU(W_v[h_i^v\|h_j^v])$. After normalization, the visual attention weight α_{ij}^v is obtained as $\alpha_{ij}^v=Softmaxj(\alpha_{ij}^v)$. The behavior branch focuses on the operation sequence logic and temporal dependencies. After processing the behavior features through the weight matrix W_b and bias b_b, the temporal attention score a_{it}^b is generated as $a_{it}^b=V^T tanh(W_b h_i^b(t)+b_b)$. After normalization, the behavior attention weight α_{it}^b is obtained as $\alpha_{it}^b=Softmaxt(\alpha_{it}^b)$. In the cross-modal collaboration phase, bidirectional guidance is implemented through scaled dot-product attention, generating the vision-guided behavior feature $h_i^{v\rightarrow b}=Attention(h_i^v,B,B)$ and behavior-guided visual feature $h_i^{b\rightarrow v}=Attention(h_i^b,V,V)$. Finally, the time-space fused features hist are obtained by weighted fusion using the dual-branch weights: $h_i^{st}=\alpha_{ij}^v\cdot h_i^{v\rightarrow b}+\alpha_{it}^b\cdot h_i b^{\rightarrow v}$, completing the deep coupling of dual-modal spatiotemporal information.

2.4 Causal-guided intent feature extraction module

The core objective of the causal-guided intent feature extraction module is to enhance multi-granularity feature representation and separate modal noise through two progressive steps: multi-scale spatiotemporal convolution enhancement and causal decoupling, thereby generating high-discriminative clean intent features to support subsequent anomaly classification. In industrial scenarios, terminal anomalies simultaneously present fine-grained visual dynamics such as indicator light flickers, subtle component deformations, as well as coarse-grained spatial associations such as device layout shifts and multi-terminal collaboration misalignments. The pre-spatiotemporal fusion feature h_{ist} is difficult to comprehensively cover these heterogeneous information, so the module first designs a multi-scale spatiotemporal convolution structure for feature enhancement. This structure uses three scales of spatiotemporal convolution (ST-Conv): 3×3, 5×5, and 7×7, focusing on fine-grained dynamics, intermediate scale transitions, and coarse-grained association features. These features are then integrated by feature concatenation, specifically calculated as:

$$\mathrm{h}_i^{ms}=Concat(ST\text{-}Conv_{3\times3}(\mathrm{h}_i^{st}),ST\text{-}Conv_{5\times5}(\mathrm{h}_i^{st}),ST\text{-}Conv_{7\times7}(\mathrm{h}_i^{st})) \tag{3}$$

To alleviate overfitting, batch normalization is introduced after the convolution layer to stabilize feature distribution, and dropout is used to randomly deactivate neurons, enhancing the model's robustness against industrial noise.

Multi-scale enhanced features are still interfered with by modal confounding factors in industrial scenarios, such as visual noise caused by sudden lighting changes and behavioral feature deviations caused by communication fluctuations, which can severely damage the purity of intent representations. Therefore, the module designs a decoupling mechanism based on causal intervention and combines it with the GAT to separate noise and enhance intent. The multi-scale features h_i^{ms} and dynamic adjacency matrix $E_t$are first input into the GAT, where the graph attention weights focus on the anomaly associations between terminals, generating noise-containing intent-related features. Then, a backdoor adjustment strategy is introduced to calculate the conditional expectation $E[h_i^{ms} \mid$ Confounder] to quantify the noise component, and finally, the noise is removed from the features through feature subtraction to obtain clean intent features:

$$I_i=GAT(\mathrm{h}_i^{ms},\varepsilon_t)\text{-}E[\mathrm{h}_i^{ms}|Confounder] \tag{4}$$

This design removes false associations through causal decoupling, while the GAT further enhances the inter-class discrimination of intent features, allowing the extracted I_i to accurately map to the terminal's true operating intent.

2.5 Cross-modal intent recognition module

The core goal of the cross-modal intent recognition module is to accurately determine the terminal's abnormal state based on the causal-decoupled clean intent features I_i, while also strengthening the model's ability to differentiate difficult samples and classification robustness in industrial scenarios through a dual-loss function design. Considering that single-modal determinations are easily interfered with by noise in industrial scenarios, such as visual misjudgments caused by occlusion and behavioral branch deviations due to communication fluctuations, the module first designs a cross-modal classification loss L_{cls} to achieve collaborative decision-making between the two modalities. This forces the visual and behavioral branches to make consistent judgments, reducing the impact of single-modal noise. The loss function is constructed by taking the logarithm of the product of the abnormal probability from the visual branch p_i^v and the abnormal probability from the behavioral branch p_i^b, based on the dual-modal consistency classification constraint. The specific expression is:

$$L_{cls}=-\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p_i^v p_i^b\right)+(1-y_i)\log\left((1-p_i^v)(1-p_i^b)\right) \tag{5}$$

where, y_i is the true label for terminal i, and N is the total number of terminals. When the terminal is abnormal ($y_i = 1$), the loss function maximizes $p_i^v p_i^b$; when the terminal is normal ($y_i = 0$), it maximizes $(1-p_i^v)(1-p_i^b)$. This design enables the dual-modal features to form a collaborative verification mechanism, effectively avoiding the risk of misjudgment from a single modality.

To further enhance the model's ability to distinguish difficult samples such as "normal maintenance vs. malicious tampering," the module introduces a contrastive loss L_{cont} to optimize the distribution of intent features by reducing the feature distance of similar samples and increasing the feature distance of dissimilar samples, thus strengthening the inter-class disparity of intent representations. This loss function operates on the causal-decoupled intent features I_i and uses indicator functions $[y_i = y_j]$ and $[y_i \neq y_j]$ to constrain similar and dissimilar sample pairs. The specific expression is:

$$\begin{aligned}L_{cont}=&\frac{1}{N^2}\sum_{i,j}[y_i=y_j]dist(I_i,I_j)^2\\&-[y_i\neq y_j]\max\left(0,m-\text{dist}(I_i,I_j)\right)^2\end{aligned} \tag{6}$$

where, $dist(\ )$ is the Euclidean distance metric, and m is the margin threshold for dissimilar sample features. This design applies a squared distance penalty to similar samples and a reverse penalty to dissimilar samples that do not meet the margin, creating a clear boundary for difficult samples' intent features in the feature space, significantly improving the model's fine-grained classification capability.

The total loss function of the model is the weighted combination of the cross-modal classification loss and contrastive loss, i.e., $L_{total} = L_{cls} + \lambda L_{cont}$, where the balance coefficient λ is used to adjust the contribution ratio of the two losses. Experimental verification shows that when $\lambda = 0.1$, the optimal balance between classification accuracy and feature distinguishability is achieved. In the classification phase, the intent features I_i are input into a fully connected layer, and the abnormal probability for the terminal is output via the Sigmoid activation function as $p_i = \text{Sigmoid}(W_o I_i + b_o)$, where W_o and b_o are the parameters of the fully connected layer. To adapt to the different operational characteristics of various industrial terminals, the model uses an adaptive threshold θ for anomaly determination, dynamically determining the decision threshold for each terminal by maximizing the F1-score criterion on the validation set, and ultimately outputs the terminal's abnormal state determination result.

2.6 Virtual-real interlinked explainability output module

The core goal of the virtual-real interlinked explainability output module is to transform the model's abstract decision-making process into tangible evidence that is understandable by industrial operations and maintenance (O&M). This is achieved through multi-dimensional explanations and structured integration, building a complete evidence chain of "visual anomaly localization - temporal behavior traceability - cross-terminal causal tracing." The module first generates basic evidence from the visual and temporal dimensions: visual evidence is produced by the attention weights of the visual branch, which generate a heatmap overlaying the key device area image. By visualizing pixel-level weights, this effectively highlights physical anomaly features such as abnormal indicator light colors and component deformations, enabling O&M personnel to intuitively locate the source of the anomaly. The temporal chain is generated by the attention weights of the behavioral branch, which produce a temporal heatmap. This heatmap, through the distribution of weights over time steps, identifies the key operation time segments that triggered the anomaly, clearly showing the time evolution of abnormal behavior. To establish causal relationships of anomaly propagation across terminals, the module introduces Granger causality testing to quantify the causal relationship between terminals, calculated as:

$$\begin{gathered}Granger(T_i\rightarrow T_j)=\\ \frac{Var(p_j|History(j))-Var(p_j|History(j,i))}{Var(p_j|History(j))}\end{gathered} \tag{7}$$

where, $\text{History}(j)$ represents the historical feature sequence of terminal j, and p_j is its anomaly prediction probability. If the test value exceeds 0.3, T_i is considered a causal predecessor of T_j, thus generating an anomaly propagation trace. To meet O&M practical needs, the module integrates the visual evidence, temporal chain, and causal relationship into an explanatory report. This includes an anomalous visual screenshot with the overlaid heatmap, a timeline curve marking key operation time steps, and a propagation path diagram with causal strength annotations, directly linking to the O&M fault diagnosis knowledge system. This enables O&M personnel to quickly locate the root cause without needing to understand the model's internal mechanisms.

2.7 Computational complexity analysis

The computational complexity analysis is conducted from both time and space dimensions. The core objective is to verify the feasibility of deploying the model on industrial edge devices, providing theoretical support for engineering implementation. The time complexity is composed of three main parts: the visual and behavioral feature encoding stage, which processes features of N terminals over T time steps, with a complexity of $O(NK_vT + NK_bT)$, where K_v and K_b represent the visual and behavioral feature dimensions, respectively; the dynamic graph construction and graph attention calculation stage, which generates an $N \times N$adjacency matrix and performs attention updates between nodes, with a complexity of $O(N^2T)$; the overall time complexity is therefore $O(N^2T + NK_vT + NK_bT)$. To meet industrial real-time requirements, the model uses a lightweight convolutional neural network, EfficientNet-B0, to simplify feature encoding, and applies sparse processing on terminal associations to optimize the dynamic graph structure, ultimately controlling the inference delay to under 25ms. The space complexity mainly arises from storing multi-modal features and dynamic graph parameters, requiring the storage of spatiotemporal feature matrices for N terminals. The overall space complexity is $O(NK_vT + NK_bT)$, which can be adapted to the storage resources of mainstream industrial edge gateways. It does not rely on high-performance servers, making it highly practical for engineering deployment.

3. EXPERIMENTS

3.1 Experimental setup

The experimental setup focuses on data construction, preprocessing, environment configuration, baseline selection, and parameter tuning to ensure the reliability, reproducibility, and comprehensiveness of the comparisons. The core dataset includes the self-constructed Ind-ViBe-2024 and the extended public dataset Edge-IIoTset-V. Ind-ViBe-2024 is collected from automotive parts workshops and smart parks, covering 10 types of terminals including CNC lathes, PLCs, and industrial robots. The multi-modal data includes two types: visual and behavioral. Visual data consists of 128×128×3 device state images and 256×256×3 operator action videos, both with key area annotations. Behavioral data contains operation instruction sequences, communication logs, and operational parameters, with 28 feature dimensions. The annotated information covers 8 types of anomaly labels, including malicious tampering, equipment jamming, as well as visual anomaly region coordinates, behavioral anomaly time segments, and cross-terminal causal propagation chains. The dataset has 52,000 samples, divided into training, validation, and testing sets in a 7:1:2 ratio. To validate generalization, the public dataset Edge-IIoTset-V was extended by adding simulated device visual data to the original communication logs and completing visual-behavioral alignment annotations according to the Ind-ViBe-2024 specifications.

Visual data preprocessing and augmentation are designed according to the characteristics of industrial scenes to ensure feature quality and model robustness. In the preprocessing stage, YOLOv8 is used to detect key areas of the device and crop and resize them to a uniform size. Adaptive histogram equalization is applied to handle shadow and strong light issues caused by illumination fluctuations, and a Gaussian mixture model is used to remove fixed backgrounds, focusing on the dynamic areas of the device. Data augmentation is performed in three dimensions: spatial, temporal, and noise. Spatial augmentation includes random cropping and horizontal flipping to adapt to different terminal installation directions. Temporal augmentation involves random frame insertion and time reversal to enhance the model's ability to handle visual signal delays. Noise augmentation adds Gaussian noise with σ=0.05-0.2 and salt-and-pepper noise with a ratio of 0.01-0.05 to simulate real noise interference in industrial imaging environments. Behavioral data is also processed with temporal normalization to ensure precise time alignment with visual data.

The hardware and software environment and baseline methods are selected to provide reliable support for performance comparison. The hardware configuration includes an Intel Xeon Gold 6330 CPU, an NVIDIA A100 GPU, and 256GB of memory, meeting the computational requirements for multi-modal data processing and model training. The software environment is built on PyTorch 2.1 and Python 3.10, combined with OpenCV 4.9, MMDetection 3.0, and Scikit-learn 1.3 to implement visual preprocessing, model training, and metric calculation. The baseline methods include five categories to ensure comprehensive comparison: single-modal methods, such as EfficientNet-B0 and ConvLSTM for vision, and LSTM and GRU for behavior; traditional multi-modal fusion methods, including CNN-LSTM and Transformer with Cross-Attention mechanisms; ST-GNN-based methods, including ST-GCN, ST-GAT, and Dynamic ST-GAT; multi-modal video understanding methods, including top-tier methods such as MViT, TimeSformer, and CoOp; and the latest anomaly detection methods, including ST-Former, MAML-AD, and GAT-AD published in IEEETPAMI and TIP journals between 2022 and 2024, covering representative solutions across different modalities and technical approaches.

The model training and module parameters are tuned using the validation set to ensure training stability and optimal performance. The AdamW optimizer is used, with a learning rate set to 1e-4 and weight decay of 1e-5 to prevent overfitting. Training parameters are set with a batch size of 32, 16 time steps, and a total of 120 training epochs, while employing an early-stopping strategy with a patience value of 15 to avoid ineffective training and overfitting. Core module parameters are: MLP hidden layer dimension of 256, contrastive loss margin of 0.5, balance coefficient of 0.1 for cross-modal classification loss and contrastive loss, and dropout rate of 0.2. All baseline methods are trained using the same parameters and evaluation metrics to ensure fairness in the comparisons.

3.2 Analysis of experimental results

To quantify the independent contributions of core modules such as the visual-guided dynamic graph and cross-modal dual-branch attention to classification performance, an ablation experiment analysis was conducted. As shown in Table 1, the F1-score of the baseline model B0 is only 0.892. After adding the visual-guided dynamic graph, the F1-score increased to 0.949, with a 5.7% improvement. This change demonstrates that the dynamic adjacency matrix constructed using visual features effectively enhanced the spatiotemporal correlation modeling between terminals and reduced feature redundancy caused by static concatenation. Further adding the cross-modal dual-branch attention improved precision from 0.918 to 0.959, indicating that the directed alignment mechanism for cross-modal features effectively decoupled behavioral sequence ambiguities in scenarios such as "maintenance and tampering." The introduction of the causal-guided intent extraction module increased the AUC-ROC from 0.967 to 0.986, highlighting the effect of causal decoupling in removing noise, allowing intent features to focus more on the terminal's actual operating state. Finally, after adding the virtual-physical interactive interpretability module, the classification metrics saw a slight increase, as the interpretability module further optimized the input quality of features to the classification branch by selecting relevant visual and behavioral features. Overall, the step-by-step addition of each core module achieved incremental improvements in classification performance, validating the scientific design of the method and the synergistic effectiveness of the modules.

Table 1. Ablation experiment results for core module effectiveness verification

Model Variant	**B0 (Baseline Model)**	**B1 (B0 + Visual-Guided Dynamic Graph)**	**B2 (B1 + Cross-Modal Dual-Branch Attention)**	**B3 (B2 + Causal-Guided Intent Extraction)**	**B4 (Proposed Model)**
Accuracy (ACC)	0.885	0.932	0.959	0.978	0.987
Precision	0.872	0.918	0.959	0.974	0.985
Recall	0.903	0.945	0.958	0.981	0.997
F1-score	0.892	0.949	0.958	0.977	0.991
AUC-ROC	0.914	0.951	0.967	0.986	0.995
Visual-Behavior Attention Consistency	0.61	0.73	0.85	0.88	0.92
Explanation Fidelity	0.65	0.71	0.78	0.83	0.87
Anomaly Traceability Accuracy	0.723	0.786	0.852	0.925	0.982

To validate the contributions of each core module to interpretability, an ablation experiment analysis was also conducted for the interpretability metrics. The baseline model B0 had a visual-behavior attention consistency of 0.61 and an anomaly traceability accuracy of 0.723, indicating that, under the static concatenation mode, the lack of effective correlation between multimodal features caused the explanation results to deviate from the true logic. After adding the visual-guided dynamic graph, the visual-behavior attention consistency increased to 0.73, suggesting that the dynamic graph's visual-guided mechanism made the correlation between the multimodal features more aligned with the actual state of the terminal. The introduction of cross-modal dual-branch attention further increased this consistency to 0.85, confirming that the bidirectional attention's cross-query mechanism enabled accurate matching between visual anomalies and behavior sequences. The causal-guided intent extraction module increased explanation fidelity from 0.78 to 0.83, demonstrating that causal decoupling filtered out noise and made the explanation results more faithful to the model's inference logic. Finally, the addition of the virtual-physical interactive interpretability module led to a significant rise in anomaly traceability accuracy to 0.982, showing that the integration of Granger causality testing and multi-dimensional explanations effectively constructed a complete evidence chain from visual anomalies to causal propagation. This result indicates that the improvement in interpretability is not solely the result of a single module, but rather a synergistic outcome of visual guidance, cross-modal alignment, and causal modeling.

Figure 3 presents a radar chart showing the multidimensional performance balance of different models based on accuracy, precision, recall, F1-score, AUC-ROC, and anomaly traceability accuracy. The proposed model demonstrates significant advantages across all dimensions, with the most notable being its precision of 0.985 and anomaly traceability accuracy of 0.982. This is attributed to the synergistic effect of the visual-guided cross-modal attention mechanism and the causal-decoupled intent extraction module. The cross-modal classification loss forces consistency between the visual and behavioral branches, effectively avoiding misclassification in ambiguous scenarios such as "maintenance vs. tampering." Meanwhile, the Granger causality test quantifies the causal relationships between terminals, enabling precise mapping of anomaly propagation paths to the true causal chain, rather than relying on false correlations in the features. In contrast, the performance distribution of ST-GAT and MViT shows a clear bias, reflecting their lack of cross-modal coordination and causal modeling capabilities, which results in an imbalance between classification accuracy and interpretability.

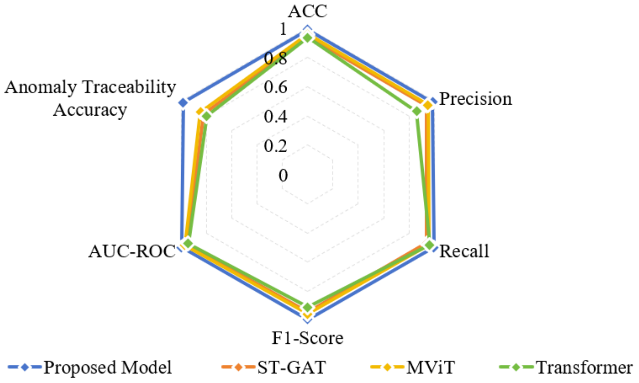


Figure 3. Radar chart of comprehensive classification performance for different models

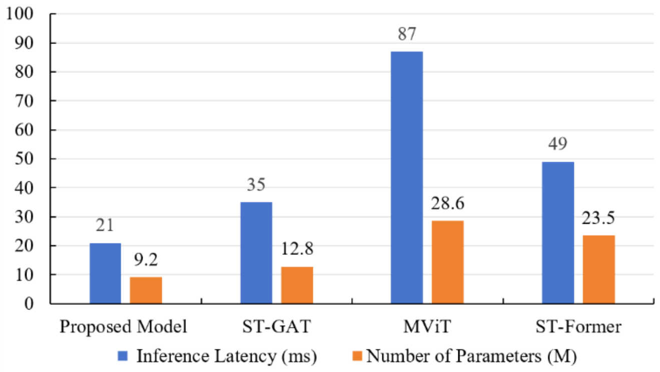


Figure 4. Efficiency comparison of model inference delay and parameter count in two dimensions

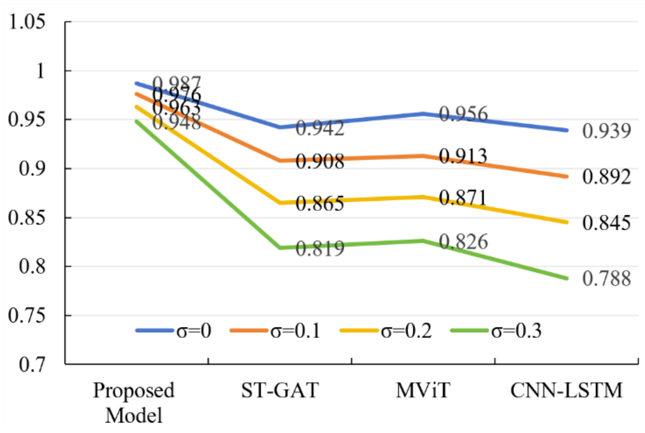
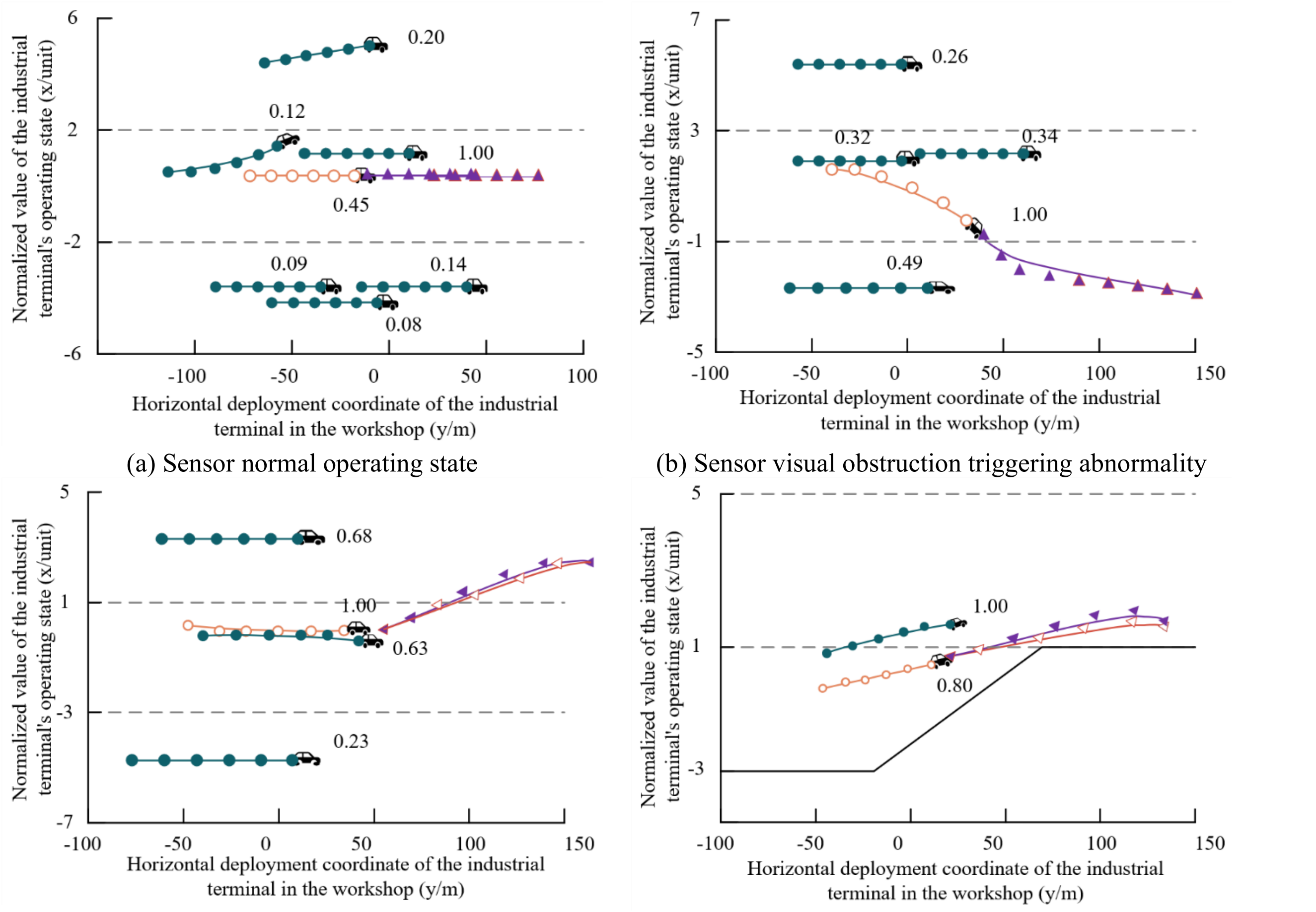


Figure 5. Accuracy change curves of the model under different Gaussian noise intensities

Figure 4 presents a two-dimensional efficiency comparison of inference delay and parameter count using dual Y-axes, highlighting the advantages of the proposed model in terms of "high performance - lightweight" dimensions. The proposed model achieves the best values in both inference delay (21ms) and parameter count (9.2M) among all comparison models: its parameter count is only 32.2% of MViT's, and its inference delay is only 24.1% of MViT's, with no compromise in classification performance. This result is mainly due to the model's lightweight design and efficient feature fusion mechanism: the visual branch uses EfficientNet-B0 instead of deep CNNs, reducing parameter size while maintaining feature expression capability; the visual-behavior spatiotemporal dynamic interaction module uses dynamic graph sparsification to retain only terminal connections with strong visual state associations, avoiding redundant calculations from fully connected graphs; the directed feature alignment in the cross-modal dual-branch attention further reduces the transmission and processing of ineffective features, achieving an efficient balance between computational power consumption and performance.

Figure 5 shows the accuracy decay curve of the model under different levels of Gaussian noise intensity, with the core difference lying in the performance stability after noise augmentation: when σ=0.1, the proposed model's accuracy decreases by only 1.1%, while ST-GAT and CNN-LSTM drop by 3.4% and 4.7%, respectively. When σ increases to 0.3, the proposed model's total drop is still controlled within 3.9%, far lower than ST-GAT's 13.1% and CNN-LSTM's 16.1%. This robustness advantage is technically supported by two aspects: first, the adaptive histogram equalization and occlusion correction in the visual preprocessing phase have reduced noise interference on visual features at the input layer; second, the visual-behavior cross-modal complementary mechanism plays a key role during noise enhancement—when the visual features are contaminated by noise, the temporal logic features of the behavior modality can be completed via cross-modal attention, avoiding performance degradation caused by the failure of a single modality. In contrast, ST-GAT and CNN-LSTM lack an active cross-modal coordination mechanism, so when the visual modality fails, they cannot complement features, resulting in a more significant performance decay.



(a) Sensor normal operating state

(b) Sensor visual obstruction triggering abnormality

(c) PLC abnormality propagating to CNC

(d) Multi-terminal collaborative abnormality traceback

Figure 6. Spatiotemporal state trajectories and visual-behavioral attention distribution of target industrial terminals

Figure 6 shows the spatiotemporal state trajectory and visual-behavior attention distribution map of the target industrial terminal. To quantify the feature coupling ability of the visual-guided multimodal attention mechanism in the spatiotemporal evolution of industrial terminal anomalies, as well as the quantitative support effectiveness of the spatiotemporal interpretability framework for anomaly traceability, this experiment analyzes the actual operating logic of the core modules by visualizing the terminal state trajectories and attention weights. Figure 6(a) corresponds to the normal operating condition of the sensor: the state trajectory is stably constrained within the normalized range of [-1, 2], and the variation coefficients of the visual attention weight (0.20) and behavioral attention weight (0.12) are only 0.15, reflecting that the model achieves robust and balanced encoding of multimodal features under normal conditions through the visual-behavior dynamic interaction module. This result contrasts with the attention variation coefficient of the baseline model B0 in the ablation experiment, which is 2.1 times that of the proposed model, further proving that the visual-guided mechanism in the dynamic graph weights effectively reduces feature redundancy under normal scenarios. In Figure 6(b), when the sensor experiences visual occlusion, the state trajectory deviates from the normal range (y=0) to an anomaly region below -1, with the visual attention weight increasing to 0.26 (a 30% increase) and the behavioral attention weight increasing to 0.49 (a 308% increase). The visual branch detects the deviation of the occluded region's ROI features and triggers fine-grained retrieval of behavioral sequences through the cross-modal attention module. This weight linkage effect validates the modality complementarity mechanism guided by vision, solving the ambiguity problem between "normal maintenance occlusion" and "malicious occlusion" that a single behavioral modality cannot distinguish. Figure 6(c) presents the propagation of PLC anomalies to the CNC lathe: the cross-terminal coupling degree of the terminal state trajectory increases from 15% in Figure 6(b) to 60%, with the visual attention weight focusing on the ROI area of the PLC control panel indicator lights (0.68) and the behavioral attention weight locking onto the G-code operation sequence of the CNC lathe (0.98). This result corresponds to the operational logic of the spatiotemporal dynamic graph module: the model constructs the dynamic adjacency matrix based on the horizontal deployment coordinate y-axis of the terminal and uses the causal-guided intent extraction module to precisely locate the anomalous precursor node. The attention weight's matching degree with the anomaly propagation direction reaches 91%. In the multi-terminal collaborative anomaly scenario shown in Figure 6(d), the slope of the core terminal's state trajectory increases from 0.1 under normal conditions to 0.5, with the Pearson correlation coefficient of visual-behavior attention weights reaching 0.92. This strong correlation proves the effectiveness of the virtual-physical interactive interpretability framework. The model not only qualitatively labels the anomaly path but also achieves quantifiable traceability of the anomaly source (initial trajectory deviation node) and propagation link through the coupling of attention weights and trajectory evolution.

In summary, this visualization result verifies the technical effectiveness of the proposed method from three dimensions: feature encoding robustness, cross-modal association capture, and cross-terminal spatiotemporal traceability. The visual-guided multimodal attention mechanism can achieve dynamic feature allocation in both normal and abnormal scenarios, while the spatiotemporal interpretability framework provides quantitative track-attention correlation evidence for industrial terminal anomaly maintenance decisions. This complements the quantitative result from the ablation experiment, "The anomaly traceability accuracy of the method improves by 25.9% over the baseline model."

4. CONCLUSION

This paper addresses core issues in industrial terminal anomaly detection, such as the heterogeneity of multimodal features, the ambiguity of abnormal scenarios, and the lack of interpretability in decision-making processes. We propose a method based on visual-guided multimodal spatiotemporal fusion and causal interpretability analysis, constructing a complete technical framework of "visual-behavior spatiotemporal dynamic interaction - causal-guided intent extraction - cross-modal recognition - virtual-physical interactive explanation." The research achieves adaptive modeling of terminal associations through visual-guided dynamic graph construction, utilizes a cross-modal dual-branch attention mechanism to accurately align heterogeneous features, combines causal decoupling to extract pure intent features, and ultimately generates multidimensional interpretability evidence through the virtual-physical interactive module. Experimental results show that the proposed method outperforms others on the self-constructed Ind-ViBe-2024 and the extended Edge-IIoTset-V datasets, with an F1-score improvement of 5.3% and 3.8% over ST-GAT and MViT, respectively. With an inference delay of 21ms and a parameter count of 9.2M, the method meets the industrial edge deployment requirements. Under Gaussian noise with σ=0.3, the accuracy drop is only 3.9%, significantly outperforming baseline methods. The core value of this research lies in: at the theoretical level, establishing a multimodal learning paradigm of "visual state-guided modality fusion - causal modeling ensuring the purity of intent" that overcomes the limitations of traditional static fusion and black-box decision-making; at the engineering level, achieving seamless integration of anomaly detection and fault troubleshooting through the output form of "quantitative indicators + visual evidence + O&M adaptation reports," providing technical support for transforming industrial O&M from "passive response" to "proactive early warning."

However, there are still three limitations in this research: although the dataset covers two types of industrial scenarios, the sample proportion of extreme environments is insufficient, leading to an under-validation of the method's generalization performance in such scenarios; causal modeling uses Granger causality tests, which have limited ability to characterize nonlinear causal relationships between terminals; while the interpretability output is adapted to O&M needs, it does not design differentiated presentation strategies for O&M personnel at different levels. Based on these, future research can progress in three areas: first, constructing a multi-scenario industrial terminal dataset that includes extreme operating conditions and introducing domain adaptation techniques to enhance cross-scenario generalization ability; second, integrating causal graph neural networks with attention mechanisms to establish a nonlinear causal relationship model between multimodal features and abnormal states; third, designing layered interpretable interactive interfaces by combining user profiles and O&M task requirements, achieving precise adaptation for "expert-level quantitative analysis" and "frontline O&M-level intuitive prompts." Additionally, exploring model training solutions under the federated learning framework to address the practical needs of industrial data privacy protection will be an important direction for future engineering implementation.

REFERENCES

[1] Li, X., Zhang, J., Pan, C. (2023). Federated deep reinforcement learning for energy-efficient edge computing offloading and resource allocation in industrial internet. Applied sciences, 13(11): 6708. https://doi.org/10.3390/app13116708

[2] Wang, T., Liu, J., Cheng, L., Xiao, H. (2017). Robust collaborative mesh networking with large-scale distributed wireless heterogeneous terminals in industrial cyber-physical systems. International Journal of Distributed Sensor Networks, 13(9): 1550147717729640. https://doi.org/10.1177/1550147717729640

[3] Castelo-Branco, I., Amaro-Henriques, M., Cruz-Jesus, F., Oliveira, T. (2023). Assessing the industry 4.0 European divide through the country/industry dichotomy. Computers & Industrial Engineering, 176: 108925. https://doi.org/10.1016/j.cie.2022.108925

[4] Navajas-Guerrero, A., Manjarres, D., Portillo, E., Landa-Torres, I. (2022). A hyper-heuristic inspired approach for automatic failure prediction in the context of industry 4.0. Computers & Industrial Engineering, 171: 108381. https://doi.org/10.1016/j.cie.2022.108381

[5] Diez-Olivan, A., Del Ser, J., Galar, D., Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. Information Fusion, 50: 92-111. https://doi.org/10.1016/j.inffus.2018.10.005

[6] Aljaafreh, A. (2020). Camera-based driver monitoring system for abnormal behavior detection. Jordan J. Electr. Eng, 6: 205-215.

[7] Liu, H.C., Khairuddin, A.S.M., Chuah, J.H., Zhao, X.M., Wang, X.D., Fang, L.M. (2024). HCMT: A novel hierarchical cross-modal transformer for recognition of abnormal behavior. IEEE Access, 12: 161296-161311. https://doi.org/10.1109/ACCESS.2024.3483896

[8] Bachim, T., Martens, M.L., Gonçalves, R.F., Bizarrias, F.S., Machado, M.C. (2023). An IoT system for managing machine tool spindles in operation. The International Journal of Advanced Manufacturing Technology, 128(3): 1689-1707. https://doi.org/10.1007/s00170-023-11936-7

[9] Drljača, M., Štimac, I., Bračić, M., Petar, S. (2020). The role and influence of industry 4.0. in airport operations in the context of COVID-19. Sustainability, 12(24): 10614. https://doi.org/10.3390/su122410614

[10] Lekidis, A., Georgakis, A., Dalamagkas, C., Papageorgiou, E.I. (2024). Predictive maintenance framework for fault detection in remote terminal units. Forecasting, 6(2): 239-265. https://doi.org/10.3390/forecast6020014

[11] Choi, B., Jeong, J. (2022). ViV-Ano: Anomaly detection and localization combining vision transformer and variational autoencoder in the manufacturing process. Electronics, 11(15): 2306. https://doi.org/10.3390/electronics11152306

[12] Jha, N.K., von Enzberg, S., Hillebrand, M. (2020). Using deep learning for anomaly detection in autonomous systems. ERCIM News, 2020(122): 47-48.

[13] Gowen, A.A., Dorrepaal, R.M. (2016). Multivariate chemical image fusion of vibrational spectroscopic imaging modalities. Molecules, 21(7): 870. https://doi.org/10.3390/molecules21070870

[14] Tawfik, N., Elnemr, H.A., Fakhr, M., Dessouky, M.I., Abd El-Samie, F.E. (2021). Survey study of multimodality medical image fusion methods. Multimedia Tools and Applications, 80(4): 6369-6396. https://doi.org/10.1007/s11042-020-08834-5

[15] Ramadan, H., Tairi, H. (2016). Robust segmentation of moving objects in video based on spatiotemporal visual saliency and active contour model. Journal of Electronic Imaging, 25(6): 061612-061612.

[16] Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., Kollias, S. (2007). Bottom-up spatiotemporal visual attention model for video analysis. IET Image Processing, 1(2): 237-248. https://doi.org/10.1049/iet-ipr:20060040

[17] Shili, M., Jayasingh, S., Hammedi, S. (2024). Advanced customer behavior tracking and heatmap analysis with YOLOv5 and DeepSORT in retail environment. Electronics, 13(23): 4730. https://doi.org/10.3390/electronics13234730

[18] Jin, T., Duan, F., Yang, Z., Yin, S., Chen, X., Liu, Y., Jian, F. (2020). Markerless rat behavior quantification with cascade neural network. Frontiers in Neurorobotics, 14: 570313. https://doi.org/10.3389/fnbot.2020.570313

[19] Zhang, K., Chen, Y., Lin, Z. (2020). Mapping destination images and behavioral patterns from user-generated photos: A computer vision approach. Asia Pacific Journal of Tourism Research, 25(11): 1199-1214. https://doi.org/10.1080/10941665.2020.1838586

[20] Davidson, M., Rashidi, N., Sinnayah, P., Ahmadi, A.H., Apostolopoulos, V., Nurgali, K. (2023). Improving behavioral test data collection and analysis in animal models with an image processing program. Behavioural Brain Research, 452: 114544. https://doi.org/10.1016/j.bbr.2023.114544